ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MANUSCRIPT-BASED THESIS PRESENTED TO
ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

BY
Miguel Angel DE LA TORRE GOMORA

ADAPTIVE MULTI-CLASSIFIER SYSTEMS FOR FACE RE-IDENTIFICATION APPLICATIONS

MONTREAL, JANUARY 26, 2015

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS:

Dr. Eric GRANGER, thesis director
Département de génie de la production automatisée, École de technologie supérieure

Dr. Robert SABOURIN, co-advisor
Département de génie de la production automatisée, École de technologie supérieure

Dr. Sylvie RATTÉ, committee president
Département de génie logiciel et des technologies de l'information, École de technologie supérieure

Dr. Jean MEUNIER, external examiner
Département d'Informatique et recherche opérationnelle, Université de Montréal

Dr. Luc DUONG, invited examiner
Département de génie logiciel et des TI, École de technologie supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON JANUARY 14, 2015

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I express my deepest gratitude to my advisers Prof. Eric Granger and Prof. Robert Sabourin, who knew how to guide my studies, with patience and tact to achieve the objectives of this phase of my life. I'm also grateful to those personages that played an important role in my education, including the collaborative work with Prof. Dmitry O. Gorodnichy, and the greatly instructive talks with Prof. Tony Wong. Since the beginning of my stay at Montréal and my studies in ETS, the operative help and personal support offered by Jorge Prieto and Estefanía Fuentes has been precise and valuable.

I would like to acknowledge the constant help, support and friendship of my friends from LIVIA: Ali Akber Dewan, Bassem R. Guendy, Donavan Prieur, Francis Quintal, Olaf Gagnon, Safa Tliba, Christophe Pagano, Jean-François Connolly, David Dubois, George Eskander, Eduardo Vellasques, Paulo R. Cavalin, Idrissa Coulibaly, Paulo V.W. Radtke, Eric Thibodeau, Dominque Rivard, Marcelo Kapp, Wael Khreich, Luana Batista, Rafael Menelau, Riadh Ksantini, Modhaffer Saidi, Jonathan Bouchard, Albert Ko, and all those supportive friends whose names escape from my limited size long term memory.

I also express my immeasurable gratitude to my family, Lidia, Sofía and José Miguel, who where always there, and whose support and company cannot be overestimated. This Thesis could not exist without their company and support. And a special mention to Sofía and José Miguel, who helped me with enthusiasm regardless their short age.

Finally, I appreciate the financial support provided by the Natural Sciences and Engineering Research Council of Canada, and the Defence Research and Development Canada Centre for Security Science Public Security Technical Program. Also appreciated is the financial support provided by the Program for the Improvement of the Professoriate (PROMEP) of the Secretariat of Public Education, Mexico, and the Mexican National Council for Science and Technology (CONACyT). And I cannot omit to mention the support of the leaders of the Universitary Center CUValles, and the University of Guadalajara, who allowed this project happen.

SYSTÈMES MULTI-CLASSIFICATEUR ADAPTATIFS POUR LA RECONNAISSANCE DE VISAGE EN APPLICATIONS DE RÉIDENTIFICATION

Miguel Angel DE LA TORRE GOMORA

RÉSUMÉ

Dans la vidéo-surveillance, les systèmes décisionnels reposent de plus en plus sur la reconnaissance de visage (RV) pour déterminer rapidement si les régions faciales capturées sur un réseau de caméras correspondent à des personnes d'intérêt. Les systèmes RV en vidéo-surveillance sont utilisés dans de nombreux scénarios, par exemple pour la détection d'individus sur la liste noire, la ré-identification de visages, et recherche et récupération. Cette thèse se concentre sur la RV vidéo-à-vidéo, où les modèles de visages sont créés avec des données de référence, puis mis à jour avec de nouvelles donées collectées dans des flux vidéo. La reconnaissance d'individus d'intérêt à partir d'images de visages capturées avec des caméras vidéo est une tâche qui représente de nombreux défis. Plus particulièrement, il est souvent supposé que l'aspect du visage des personnes cibles ne change pas au fil du temps, ainsi que les proportions des visages capturés pour des individus cibles et non-cibles sont équivalentes, connues *a priori* et fixes. Cependant, de nombreuses variations peuvent se manifester dans les conditions d'observation, par exemple l'éclairage, le brouillage, la résolution, l'expression, la pose et l'interopérabilité avec la caméra. De plus, les modèles de visages utilisés pour calculer des correspondances ne sont généralement pas représentatifs car désignés *a priori*, avec une quantité limitée d'échantillons de référence qui sont collectés et étiquetés à un coût élevé. Enfin, les proportions des individus cibles et non-cibles changent continuellement durant le fonctionnement du système.

Dans la littérature, des systèmes adaptatifs multi-classificateur (en anglais, multiple classifier systems, MCS) ont été utilisés avec succès pour la RV vidéo-à-video, où les modèles de visages de chaque individu cible sont générés en utilisant un ensemble de classificateurs à 2-classes (entraînés avec des échantillons cibles et non-cibles). Des approches plus récentes utilisent des ensembles de classificateurs Fuzzy ARTMAP à deux classes, entraîné avec une stratégie DPSO (*dynamic particle swarm optimization*) pour générer un groupement de classificateurs dont les paramètres sont optimisés, ainsi que la combinaison Booléenne pour la fusion de leur réponses dans l'espace ROC (*Receiver Operating Characteristics*). Des ensembles actifs de classificateurs sensibles au biais ont été récemment proposés, pour adapter la fonction de fusion d'un ensemble selon le débalancement des classes mesuré sur des données opérationnelles. Ces approches estiment les proportions cibles contre non-cibles périodiquement au cours des opérations. La fusion des ensembles de classificateurs est ensuite adaptée à ce débalancement des classes. Finalement, le suivi du visage peut être utilisé pour regrouper les réponses du système liées à une trajectoire du visage (captures du visage d'une seule personne dans la scène) pour une reconnaissance spatio-temporelle robuste, ainsi que pour mettre à jour les modèles du visage au cours du temps à l'aide des données opérationnelles.

Dans cette thèse, des nouvelles techniques sont proposées pour adapter les modèles de visages pour des individus enrôlés dans un système de RV vidéo-à-vidéo. L'utilisation de stratégies d'auto-mise à jour basées sur l'utilisation de trajectoires est proposée pour mettre à jour le système, en considérant les changements brusques et progressifs dans l'environnement de classification. Ensuite, des classificateurs adaptatifs sensibles au biais sont proposés pour l'adaptation du système au débalancement des classes lors de la phase opérationnelle.

Dans le chapitre 2, un cadre adaptatif est proposé pour l'apprentissage partiellement supervisé des modèles de visages au fil du temps en fonction des trajectoires capturées. Lors des opérations, des informations recueillies à l'aide d'un suivi de visages et des ensembles de classificateurs spécifiques à l'individu sont intégrés pour la reconnaissance spatio-temporelle robuste et l'auto-mise à jour des modèles du visage. Le suiveur définit une trajectoire de visage pour chaque personne qui apparaît dans une vidéo. La reconnaissance d'un individu cible passe si les prédictions positives accumulées d'une trajectoire dépassent un seuil de détection pour un ensemble. Lorsque le nombre de prédictions positives dépassent un seuil de mise à jour, tous les échantillons du visage de la cible de la trajectoire sont combinés avec des échantillons non-cibles (choisi parmi le modèle cohorte et le modèle universel) pour mettre à jour le modèle du visage correspondant. Une stratégie *learn-and-combine* est utilisée pour éviter la corruption de la connaissance lors de l'auto-mise à jour des ensembles. En outre, une stratégie de gestion de la mémoire basée sur la divergence Kullback-Leibler est proposée pour ordonner et sélectionner des échantillons de référence cible et non-cible les plus pertinents. Ensuite, les échantillons choisis sont stockés dans la mémoire alors que les ensembles évoluent. Pour une preuve de concept, le système proposé a été validé avec des données synthétiques et vidéos de la base de données *Face in Action*, émulant un scénario de vérification passeport. Les résultats mettent en valeur la réponse des systèmes proposés à des changements graduels et brusques dans l'apparence des visages des individus, tels que l'on trouve dans la vidéo-surveillance, dans des conditions semi-contrôlées ou non contrôlées de capture. Initialement, les trajectoires capturées à partir de vidéos de référence sont utilisées pour l'apprentissage supervisé des ensembles. Ensuite, des vidéos de plusieurs scénarios opérationnels ont été présentés au système, qui a été automatiquement mis-à-jour avec des trajectoires de haut niveau de confiance. Une analyse des résultats image par image avec des données réelles montre que l'approche proposée surpasse les systèmes de référence qui ne s'adaptent pas aux nouvelles trajectoires. De plus, le système proposé offre des performances comparables à des systèmes idéaux qui s'adaptent à toutes les trajectoires cibles concernées, à travers l'apprentissage supervisé. Une analyse par individu révèle la présence d'individus particuliers, pour lesquels les ensembles automatiquement mis à jour avec les trajectoires de visages sans étiquette présentent un avantage considérable. Enfin, une analyse au niveau des trajectoires révèle que le système proposé permet une RV vidéo-à-vidéo robuste.

Dans le chapitre 3, une extension et une mise en oeuvre particulière du système de RV spatio-temporelle utilisant des ensembles est proposée, et il est caractérisé en scénarios avec des changements progressifs et brusques dans l'environnement de classification. L'analyse des résultats image par image montrent que le système proposé permet d'augmenter la précision AUC (surface sous la courbe ROC) d'environ 3 % dans les scénarios avec des changements

brusques, et d'environ 5 % dans les scénarios avec des changements graduels. Une analyse par sujet révèle les limitations de la reconnaissance de visage avec des variations de pose, affectant plus de façon significative les individus de type agneaux et chèvre. Par rapport à des approches de fusion spatio-temporelle de référence, les résultats montrent que l'approche proposé présente une meilleure capacité de discrimination.

Dans le chapitre 4, des ensembles adaptatifs sont proposés pour combiner des classificateurs entraînés avec des niveaux de débalancement et complexité variables pour améliorer la performance dans la RV vidéo-à-video. Lors des opérations, le niveau de débalancement est périodiquement estimé à partir des trajectoires d'entrée utilisant la méthode de quantification HDx, et des représentations d'histogrammes pré-calculés de la distribution des données débalancées. Les réponses des ensembles sont accumulées pour la reconnaissance vidéo-à-vidéo sensible au débalancement. Les résultats sur les données synthétiques montrent qu'en utilisant l'approche proposée, on observe une amélioration significative de la performance. Les résultats sur des données réelles montrent que la méthode proposée surpasse la performance des techniques de référence dans des environnements de surveillance vidéo.

Mots clés: Systèmes multi-classificateur, reconnaissance adaptatif de visages, apprentissage semi-supervisé, combinaison sensible au biais, débalancement de classes

ADAPTIVE MULTI-CLASSIFIER SYSTEMS FOR FACE RE-IDENTIFICATION APPLICATIONS

Miguel Angel DE LA TORRE GOMORA

ABSTRACT

In video surveillance, decision support systems rely more and more on face recognition (FR) to rapidly determine if facial regions captured over a network of cameras correspond to individuals of interest. Systems for FR in video surveillance are applied in a range of scenarios, for instance in watchlist screening, face re-identification, and search and retrieval. The focus of this Thesis is video-to-video FR, as found in face re-identification applications, where facial models are designed on reference data, and update is archived on operational captures from video streams. Several challenges emerge from the task of recognizing individuals of interest from faces captured with video cameras. Most notably, it is often assumed that the facial appearance of target individuals do not change over time, and the proportions of faces captured for target and non-target individuals are balanced, known *a priori* and remain fixed. However, faces captured during operations vary due to several factors, including illumination, blur, resolution, pose expression, and camera interoperability. In addition, facial models used matching are commonly not representative since they are designed *a priori*, with a limited amount of reference samples that are collected and labeled at a high cost. Finally, the proportions of target and non-target individuals continuously change during operations.

In literature, adaptive multiple classifier systems (MCSs) have been successfully applied to video-to-video FR, where the facial model for each target individual is designed using an ensemble of 2-class classifiers (trained using target vs. non-target reference samples). Recent approaches employ ensembles of 2-class Fuzzy ARTMAP classifiers, with a DPSO strategy to generate a pool of classifiers with optimized hyperparameters, and Boolean combination to merge their responses in the ROC space. Besides, the skew-sensitive ensembles were recently proposed to adapt the fusion function of an ensemble according to class imbalance measured on operational data. These active approaches estimate target vs. non-target proportions periodically during operations distance, and the fusion of classifier ensembles are adapted to such imbalance. Finally, face tracking can be used to regroup the system responses linked to a facial trajectory (facial captures from a single person in the scene) for robust spatio-temporal recognition, and to update facial models over time using operational data.

In this Thesis, new techniques are proposed to adapt the facial models for individuals enrolled to a video-to-video FR system. Trajectory-based self-updating is proposed to update the system, considering gradual and abrupt changes in the classification environment. Then, skew-sensitive ensembles are proposed to adapt the system to the operational imbalance.

In Chapter 2, an adaptive framework is proposed for partially-supervised learning of facial models over time based on facial trajectories. During operations, information from a face

tracker and individual-specific ensembles is integrated for robust spatio-temporal recognition and for self-update of facial models. The tracker defines a facial trajectory for each individual in video. Recognition of a target individual is done if the positive predictions accumulated along a trajectory surpass a detection threshold for an ensemble. If the accumulated positive predictions surpass a higher update threshold, then all target face samples from the trajectory are combined with non-target samples (selected from the cohort and universal models) to update the corresponding facial model. A learn-and-combine strategy is employed to avoid knowledge corruption during self-update of ensembles. In addition, a memory management strategy based on Kullback-Leibler divergence is proposed to rank and select the most relevant target and non-target reference samples to be stored in memory as the ensembles evolves. The proposed system was validated with synthetic data and real videos from Face in Action dataset, emulating a passport checking scenario. Initially, enrollment trajectories were used for supervised learning of ensembles, and videos from three capture sessions were presented to the system for FR and self-update. Transaction-level analysis shows that the proposed approach outperforms baseline systems that do not adapt to new trajectories, and provides comparable performance to ideal systems that adapt to all relevant target trajectories, through supervised learning. Subject-level analysis reveals the existence of individuals for which self-updated ensembles provide a considerable benefit. Trajectory-level analysis indicates that the proposed system allows for robust spatio-temporal video-to-video FR.

In Chapter 3, an extension and a particular implementation of the ensemble-based system for spatio-temporal FR is proposed, and is characterized in scenarios with gradual and abrupt changes in the classification environment. Transaction-level results show that the proposed system allows to increase AUC accuracy by about 3% in scenarios with abrupt changes, and by about 5% in scenarios with gradual changes. Subject-based analysis reveals the difficulties of FR with different poses, affecting more significantly the lamb- and goat-like individuals. Compared to reference spatio-temporal fusion approaches, the proposed accumulation scheme produces the highest discrimination.

In Chapter 4, adaptive skew-sensitive ensembles are proposed to combine classifiers trained by selecting data with varying levels of imbalance and complexity, to sustain a high level the performance for video-to-video FR. During operations, the level of imbalance is periodically estimated from the input trajectories using the HDx quantification method, and pre-computed histogram representations of imbalanced data distributions. Ensemble scores are accumulated of trajectories for robust skew-sensitive spatio-temporal recognition. Results on synthetic data show that adapting the fusion function with the proposed approach can significantly improve performance. Results on real data show that the proposed method can outperform reference techniques in imbalanced video surveillance environments.

Keywords: Multiple Classifier Systems, Adaptive Face Recognition, Semi-Supervised Learning, Skew-Sensitive Combination, Class Imbalance

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABREVIATIONS

AMS	Average Margin Sampling
AS	Average Surprise
AUC	Area Under the ROC Curve
BC	Boolean Combination
CAMSHIFT	Continuously Adaptive Mean Shift Tracking Algorithm
CM	Cohort Model
CMU-FIA	Carnegie Mellon University-Faces In Action Face Database
CNN	Condensed Nearest Neighbor
DPSO	Dynamic Particle Swarm Optimization
EoC	Ensemble of Classifiers
EoD	Ensemble of Detectors (1- or 2-class classifiers)
FAM	Fuzzy ARTMAP Classifier
FN	False Negatives
fnr	False Negative Rate
FP	False Positives
fpr	False Positive Rate
FR	Face Recognition
FRiVS	FR in Video Surveillance
HD	Hellinger Distance

HDx	Quantification method based on Hellinger distance at feature level
HDy	Quantification method based on Hellinger distance at score level
hne	Higher Negative Envelope
IBC	Iterative Boolean Combination
IVT	Incremental Visual Tracker
KL	Kullback-Leibler divergence
*k*NN	*k* Nearest Neighbor classifier
LTM	Long Term Memory
MCS	Multiple Classifier System
MSE	Mean Squared Error
MS-LBP	Multi Scale- Local Binary Patterns
OSS	One Sided Selection
*p*AUC (5%)	Partial Area Under the ROC Curve for $0 \leq fpr \leq 0.05$
pac	Positive Accumulation Curve
PCA	Principal Component Analysis
PFAM	Probabilistic Fuzzy ARMAP classifier
PROC	Precision-Recall Receiver Operating Characteristics
ROC	Receiver Operating Characteristics
ROCCH	ROC Convex Hull
ROI	Region Of Interest

SSBC	Skew-Sensitive Boolean Combination
TCM-kNN	Transductive Confidence Machine - k Nearest Neighbor
TN	True Negatives
tnr	True Negative Rate
TP	True Positives
tpr	True Positive Rate
UM	Universal Model
VE	Vote Entropy
VS	Video Surveillance

LISTE OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathbf{a}$	Feature vector extracted from a facial region of interest
c_m	Classifier member of the EoD_k, $m = 1...M$
d_k^*	Decision produced by EoD_k given an input $\mathbf{a}$
d_m	Decision produced by c_m after applying γ_m to s_m^+
D^c	Validation data used to estimate the combination ROC curve; D^{c+} contains only positives
D^e	Validation data used to stop the training epochs; D^{e+} contains only positives
D^f	Validation data used for fitness estimation and hyperparameter optimization; D^{f+} contains only positives
D_L	Labeled dataset used for facial model design
D	Unlabeled adaptation set, $D = \{d_1, ..., d_L\}$
D_1, D_2, D_3	Unlabeled adaptation sets from capture sessions 1, 2 and 3 respectively
D^s	Validation data used to select the operations point; D^{s+} contains only positives
D^t	Training data; D^{t+} contains only positives
EoD_k	Ensemble of detectors dedicated to individual k
$\mathscr{F}_k$	Fusion function for the EoD_k
$\mathscr{F}'_k$	Updated fusion function in EoD'_k
$\mathscr{G}$	Gallery of templates used in template matching and original self- and co-update

$\mathscr{G}'$	Updated gallery of templates
Γ_k	Additive constant for the estimation of the update threshold γ_k^u
γ_k^d	Detection threshold for the EoD_k
γ_m	Decision threshold for the classifier c_m in the EoD_k
γ^T	Threshold for tracking quality
γ_k^u	Update threshold for the EoD_k
k	Label for an individual of interest, $k = 1, ..., K$
K	The number of individuals of interest enrolled to the system
λ_k	The size of the individual specific LTM_k
L	Number of templates in D
LTM_k	The long term memory used for module k
M	Number of base classifiers in the EoD_k
N	Number of templates in the gallery $\mathscr{G}$
$\mathscr{P}_k$	Pool of classifiers in the EoD_k
$\mathscr{P}'_k$	Updated pool of classifiers in EoD'_k
Q_T	Quality of a trajectory at a given state (frame)
s_m^+	Classification score for the target individual from classifier c_m
T	An unlabeled trajectory
T_k	A labeled trajectory from individual k
$t_n \in \mathscr{G}$	A template in the gallery

INTRODUCTION

Video-based face recognition (FR) is employed more and more to assist operators of intelligent video surveillance (VS) systems in industry and public sectors, due in large part to the low cost camera technologies and the advances in the areas of biometrics, pattern recognition and computer vision. Decision support systems are employed in crowded scenes (airports, shopping centers, stadiums, etc.), where an human operator monitors live or archived videos to analyze a scene (Hampapur *et al.*, 2005). VS systems perform a growing number of functions, ranging from real time recognition and video footage analysis to fusion of video data from different sources (Gouaillier, 2009). FR in VS (FRiVS) can be employed in a range of still-to-video (as found in, e.g., watchlist screening) and video-to-video (as found in, e.g., face re-identification) applications. In still-to-video FR, a gallery of still images is employed in the construction of facial models, whereas in video-to-video FR facial models are designed from video streams.

Of special interest in this Thesis is the automatic detection of a target individual of interest enrolled to a video-to-video FR system. In this human-centeric scenario, live or archived videos are analyzed, and the operator receives an alarm if it detects the presence of a target individual enrolled to the system. Due to the high amount of non-target individuals appearing in crowded scenes, avoiding false alarms while maintaining a high detection rate is challenging for such a system. The design of a FR system for real world applications raises many challenges.

Problem Statement

FR systems employed in VS face numerous problems that are related to the time and spatial variations in the real world capture conditions. For instance, the natural ageing of people induce gradual changes in the facial appearance of enrolled individuals after enrollment. Besides, variations in capture conditions like the position of the camera, lighting and pose induce abrupt changes in the classification environment. In addition, facial models are designed *a priori* with a limited amount of reference faces that are often captured under controlled conditions at enrollment time, and therefore loose their representativeness over time. The matching process

is also challenging due to changes in camera interoperability issues. The performance of a system for video-to-video FR is significantly degraded due to these factors.

Several classification systems have been proposed that can be employed for face matching in VS applications (De-la Torre *et al.*, 2012b; Li and Wechsler, 2005; Pagano *et al.*, 2012; Polikar *et al.*, 2001). Recent approaches take advantage of modular architectures with one ensemble of 2-class classifiers to design the facial model of each target individual (trained using target vs. non-target reference samples) Pagano *et al.* (2012). These modular strategies reduce the complexity of the problem faced by multi-class classifiers to find multiple decision frontiers, and add the robustness of ensemble techniques.

Adaptive multiple classifier systems (MCS) capable of incremental learning allow to update the facial models with new reference facial captures (Polikar *et al.*, 2001; Connolly *et al.*, 2010a; De-la Torre *et al.*, 2012a). For example, systems like (De-la Torre *et al.*, 2012a,b) allow to design a facial model for each target individual using an adaptive ensemble of 2-class classifiers. However, the requirement of manual acquisition and labeling of the new reference data is costly or unfeasible in practice.

The proposed strategy to address this problem consists in a system that is initially designed with reference samples, and is capable of learning highly confident operational data through self-update, improving this way the representativeness of the facial models. However, techniques in literature are not adapted for video-to-video FR and face re-identification applications. Adaptive biometric systems have been proposed to incorporate new reference samples based on semi-supervised learning schemes (Rattani, 2010; Marcialis *et al.*, 2008; Poh *et al.*, 2009). Self-update strategies allow to reduce or eliminate this labeling cost at expenses of some false updates, affecting a trade-off between self-adaptation and accuracy of facial models.

The effects of the differences in class proportions (imbalance) the performance of classifiers have been widely studied in pattern recognition literature Guo *et al.* (2008); Landgrebe *et al.* (2006); Forman (2006); Lopez *et al.* (2013), and several ensemble-based methods to train ensembles on imbalanced data have been proposed Galar *et al.* (2011). Algorithms designed

for environments with data distributions that change over time can be categorized according to the use of a mechanism to detect concept drift or change Ditzler and Polikar (2013). Active approaches seek explicitly to determine whether and when a change has occurred in the class proportions before taking a corrective action Radtke *et al.* (2013a,b); Ditzler and Polikar (2013). Conversely, passive approaches assume that a change may occur at any time, or is continuously occurring, and hence the ensembles are updated every time new data becomes available Ditzler and Polikar (2013); Oh *et al.* (2011). The advantage of active approaches mainly consists in the avoidance of unnecessary updates. However, they are prone to both false positive and false negative drift detections, with the respective false updates and false no-updates. Passive approaches avoid some of these problems at an increased computational cost due to the constant update.

A representative example of active approaches for changing imbalances is the skew-sensitive Boolean combination (SSBC) that continuously estimates the class proportions using the Hellinger distance between histogram representations of operational and validation samples Radtke *et al.* (2013b). Every time the operational imbalance changes, SSBC selects one of the pre-calculated fusion functions that correspond to a set of prefixed imbalances. However, the limited number of validation imbalance levels that can be used to approximate the imbalance in operations is a limiting factor for the estimation of operational imbalance. Rather than selecting the closest imbalanced histogram representations, more sophisticated estimation methods may be employed for accurate estimation of the class proportions. Moreover, although it is scarcely exploited, the abundant non-target samples in video surveillance allow to produce training sets with different complexities and imbalances, and use them to generate diverse pools. A specialized combination and selection scheme of these diversified pools may lead to robust ensembles, considering both the different levels of complexity and imbalance Lopez *et al.* (2013).

This discussion raises various research questions that require to be addressed. For instance, what kind of architecture would allow for facial models that provide the best performance for each of the individuals enrolled to the system? Given the abundant videos from non-target individuals available for system design, what is a good strategy to select representative non-target

samples to train an individual specific ensemble, and yet avoid a bias toward the non-target class? Since adaptive MCS employ a long term memory (LTM) to avoid knowledge corruption while learning incrementally, what is an effective strategy to avoid running out of resources after several updates? Which is a good strategy to combine spatial and temporal informations from videos? How the system operates in scenarios with gradual and abrupt changes in the distribution of faces in the feature space? Given the abundant non-target samples, can these samples be employed to train the ensembles that perform better under imbalanced conditions? And finally, how individual-specific ensembles can be efficiently adapted under abrupt and gradual environmental changes and inconstant proportions of target and non-target individuals?

Objective and contributions

In this Thesis, a new framework for adaptive MCSs is proposed for partially-supervised learning of facial models over time based on facial trajectories. This framework is designed to implement systems for video-to-video FR, as needed for face re-identification applications, where gradual or abrupt environmental changes occur over time. In Bayesian decision theory, these changes correspond to changes in the probability density function of the faces (e.g. appearance of the face), or the prior probabilities (class proportions). The main contribution of this Thesis includes the proposal of an adaptive MCS for video-to-video FR for video surveillance, capable of spatio-temporal recognition and self-updating based on highly confident facial trajectories captured in scene. The system is also capable of adapting the fusion function of individual-specific classifiers to the operational imbalance in video-to-video FR. This contribution is divided into three parts.

The first part (**Chapter 2**) consists in the proposal of a whole framework for partially-supervised learning of facial models, which adapts over time based on operational face trajectories. The proposed framework consists of a segmentation module for face detection, a face tracker, a individual-specific modular classification system, a decision fusion system, a design/update system, and a sampling selection system. On the whole, it provides the mechanisms for the

design and self-update of individual-specific facial models based on a modular classification system with one adaptive ensemble of detectors (EoD) per individual of interest. During operations, tracking IDs are combined with the responses from individual-specific ensembles for robust spatio-temporal recognition and for self-update of facial models. Trajectories are formed by regrouping facial regions with the same tracking ID (provided by the tracker), ensuring that all belong to the same individual that appears in a video.

Recognition of a target individual is achieved when the positive predictions accumulated along a trajectory surpass an individual-specific detection threshold. If the accumulated positive ensemble predictions surpass a higher update threshold, then all target face samples from the trajectory are combined with non-target samples to update the corresponding facial model. The most representative non-target samples for training and validation are selected from the cohort and universal models employing condensed nearest neighbor (CNN) selection, and a learn-and-combine strategy is employed to avoid knowledge corruption during self-update of ensembles, and Boolean combination (BC) is used to combine classifiers. In addition, a selection strategy for memory management based on Kullback-Leibler divergence is proposed to rank and select the most relevant target and non-target reference samples to be stored in memory as the ensembles evolves.

In **Chapter 3**, a particular implementation of the adaptive MCS is proposed to maintain and update independent detectors, using ensembles of 2-class PFAM classifiers per individual to discriminate between the target and non-target individuals. The one-sided selection (OSS) strategy is employed to select non-target training samples (replacing the original CNN), and the PFAM classifiers are generated using a DPSO training strategy. A learn and combine strategy is employed for adaptation of facial models, avoiding this way the corruption of knowledge. Iterative boolean combination (IBC) is employed to dynamically select individual thresholds and combination functions in ROC space. This implementation of the system was characterized using the CMU-FIA database, emulating a scenario with gradual changes (e.g. aging), and abrupt changes (e.g. pose). For a wide picture of the system operation under real world con-

ditions, a global evaluation was performed using a three-levels analysis: transaction-, subject- and trajectory-based.

Finally, in **Chapter 4**, adaptive skew-sensitive ensembles are proposed to adapt the system to the continuously changing operational imbalance. In the proposed active approach, the operational imbalance is approximated with HDx quantification, showing a lower mean squared error when compared to other techniques (Gonzalez-Castro *et al.*, 2013). And the generation of base classifiers take advantage of the availability of abundant non-target samples to train ensemble members on different levels of complexity and imbalance (Lopez *et al.*, 2013).

For proof-of-concept, the proposed system was validated with synthetic data and real videos from Face in Action dataset, emulating a passport checking scenario. The analysis with real data is divided in three levels. Transaction-based analysis considers the response of the system to each captured facial region, and performance measures are drawn in the ROC and PROC spaces. Subject-based analysis evaluates the performance of the system for each target individual, employing the categorization provided by the Doddington zoo taxonomy. Finally, trajectory-based analysis allows to evaluate the overall performance of the system involving all the modules (segmentation, tracking, classification and decision fusion).

Structure of the Thesis

This Thesis is organized into four chapters that describe the different parts of the steps followed through the advance of the research process (see Figure 0.1). Chapter 1 presents a survey of the most recent advances of FR in video and adaptive biometrics, as well as the pattern recognition concepts used for design and evaluation of the system after each experimentation.

In Chapter 2 a framework for partially-supervised learning of facial models over time based on facial trajectories is proposed and described. Chapter 2 was published as a special issue article in the journal Information Fusion from Elsevier (De-la Torre *et al.*, 2014a).

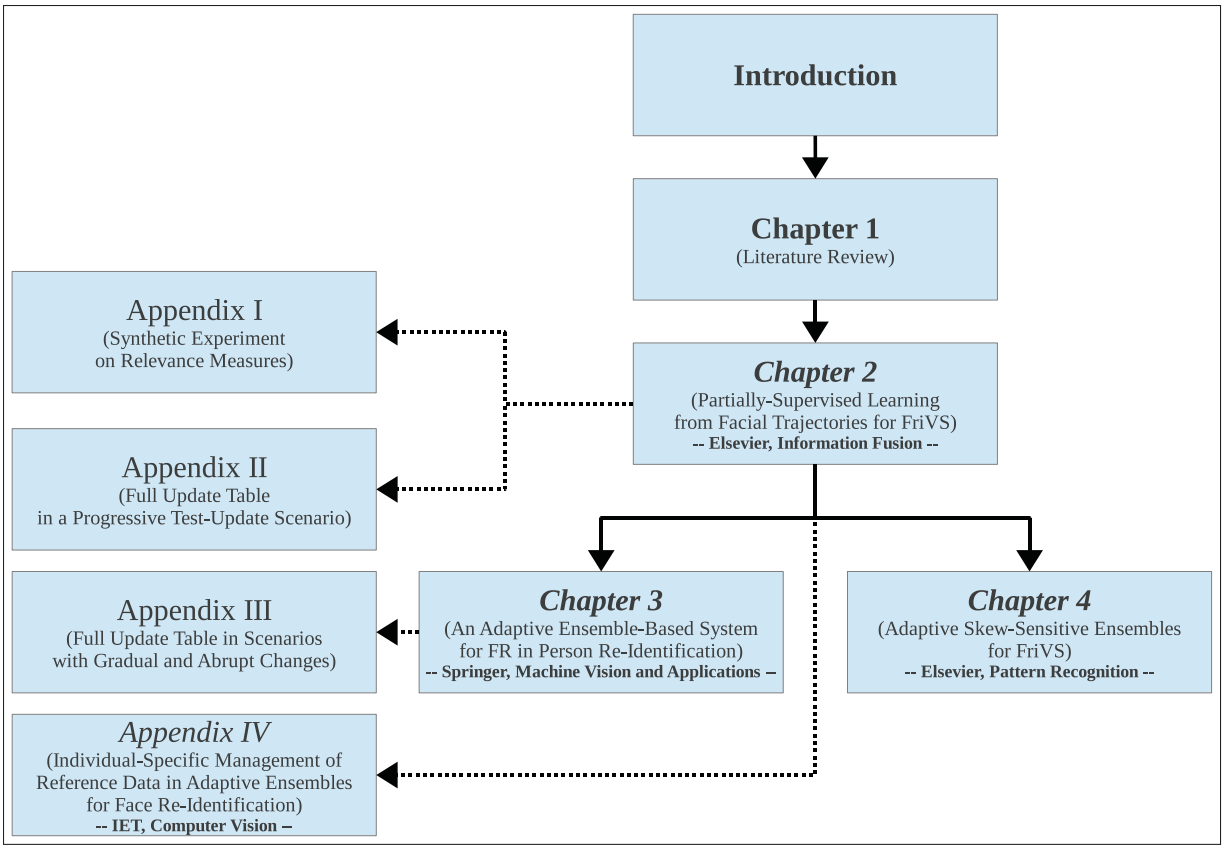


Figure 0.1 Structure of the Thesis. Solid arrows indicate the sequence of the chapters, whereas dotted arrows indicate the relationship between a chapter and the appendixes. Underlined titles in the boxes indicate that the material in the chapter (or appendix) has been submitted to a journal for publication

Appendix IV describes an individual-specific strategy for the management of reference samples stored in a long term memory, suitable to be employed in the framework proposed in Chapter 2. The content of Appendix IV corresponds to an extended version of a paper presented in the international conference on imaging for crime prevention and detection (De-la Torre *et al.*, 2013), and was submitted to the journal IET-Computer Vision after invitation.

Chapter 3 describes the proposed implementation of the ensemble-based system for spatio-temporal FR and the selection of parameters for gradual and abrupt changes in the classification environment. Chapter 3 was accepted with revision to be published in the journal Machine Vision and Applications from Springer (De-la Torre *et al.*, 2014b).

In Chapter 4, skew-sensitive ensembles have been proposed for adaptive skew-sensitive FR in video surveillance, employing a strategy that takes advantage of various levels of complexity and imbalance to design ensembles of classifiers. Chapter 4 was submitted to the journal Pattern Recognition from Elsevier.

CHAPTER 1

A REVIEW OF TECHNIQUES FOR ADAPTIVE FACE RECOGNITION IN VIDEO SURVEILLANCE

Intelligent video surveillance systems that employ face recognition (FR) for decision support are important in many private, but mostly public sector applications. The extensive use of FR systems is due in part to the universality of the human face as a biometric trait that can be covertly captured, the availability of low cost cameras, and to advances in biometrics, pattern recognition and image/video processing. These systems are being considered for video surveillance in crowded scenes (airports, shopping centers, stadiums, etc.) In these scenes, an operator observes the scene through surveillance cameras and monitor who or what is in scene (Hampapur *et al.*, 2005). Although many decision support systems exist, there are still many functions to be developed or improved. These areas of opportunity for researchers range from the real time recognition to fusion of video data from different sources, passing through the design of compact biometric models and the preservation of performance over time (Gouaillier, 2009; Ahmad *et al.*, 2008). Of special interest in this Thesis is the automatic detection of individuals of interest enrolled to a system, based on the appearance of their face, and the preservation of system's performance regardless of variations over time of a target individual's appearance.

In the human-centeric scenario assumed in this thesis, an operator monitors the surveilled place employing an intelligent video surveillance system capable of video-to-video FR (for applications like, e.g. real-time monitoring or search and retrieval from video archive.) The system generates a set of alarms on each of the individuals of interest enrolled to the system, and the operator must confirm the detection –that the individual detected by the system truly corresponds to an individual of interest. In this scenario, facial models are designed considering spatial and temporal information from video streams for video-to-video FR.

In this Chapter, a literature review was conducted in the different areas related to the most recent advances on systems for video-to-video FR for face re-identification. The Chapter sum-

marizes the most up-to-date academic systems and technologies for FRiVS, semi-supervised learning, adaptive biometrics, incremental and on-line learning of classifiers, and adaptive skew-sensitive ensembles including passive and active approaches.

1.1 Face Recognition in Video-Surveillance

Video technologies have been widely investigated over the last years (Zhao *et al.*, 2003; Wang *et al.*, 2009; Matta and Dugelay, 2009). Challenges addressed by video based FR systems are being investigated from diverse research areas, including computer vision, pattern recognition and perception. Figure 1.1 depicts a general biometric system for video to video FR, where one or several cameras capture the real world scene over time, and the system responds according to its particular functionality. According to Figure 1.1, the video frames feed a segmentation module that detects and isolates the facial regions of interest (ROIs) used for tracking and classification. The tracking system follows the facial ROI across frames, whereas the classification system compares feature representations of the input ROIs against facial models stored in a biometric database. Then, tracking IDs and classification scores are combined for enhanced spatio-temporal recognition.

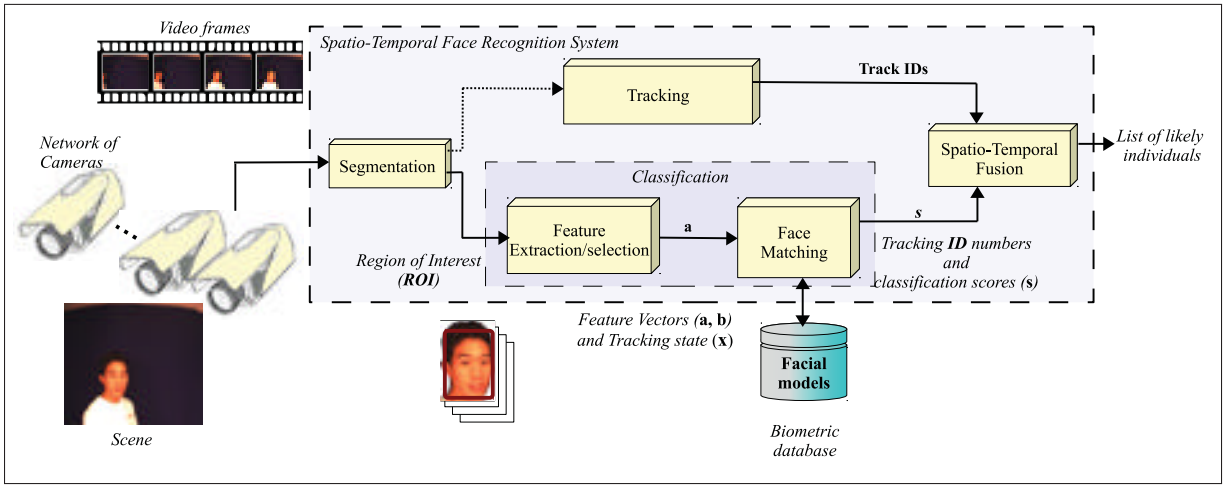


Figure 1.1 A general biometric system for FRiVS

Segmentation consists on isolating and retrieving the facial region of interest (ROI) (position and the pixels) in the input frame(s). The wide range of techniques appearing in literature are generally categorized into four types. *Knowledge based* methods work with predefined rules based on human knowledge to determine if an image is a face. *Feature invariant* approaches find face structure related features, which are robust to pose and lighting conditions. *Template matching* methods employ pre stored face templates to decide if an image is a face. *Appearance-based* methods use learning strategies to produce face models from a set of training samples. The general trend indicates that appearance-based methods produce superior performance than others (Yang *et al.*, 2002), and the most representative example of an appearance-based method is the so called Viola-Jones algorithm (Viola and Jones, 2004). In that work authors represent images using AdaBoost to create a classifier that selects critical visual features, and propose a "cascade" classifier combination method to extract faces. Since the Viola-Jones algorithm, boosting based face detection schemes have evolved as de-facto standard of face detection in real world applications (Zhang and Zhang, 2010).

The feature extraction module extracts and selects the most discriminant measurable characteristics from the ROIs to form a feature vector. A good feature extractor would yield a representation that facilitates the task of classification. *Holistic*, *feature-based (structural)* and *hybrid* matching methods require then compatible features. Of special interest are the holistic features, which use the whole face (ROI) as the raw input to the FR system. Common feature extraction techniques of this type are the eigenfaces based on principal component analysis (PCA) (Turk and Pentland, 1991), Fisherfaces that take advantage of linear discriminant analysis (LDA) (Belhumeur *et al.*, 1997; Martinez and Kak, 2001) or a combination of both LDA and PCA (Marcialis and Roli, 2002). Independent component analysis has been also proposed as a generalization of PCA (Bartlett *et al.*, 2002), with the advantages that the transformation matrix can be estimated with limited/unlabeled data, both local and global features are considered and higher order statistics between pixels/images are exploited. Variants of PCA which take advantage of information of the two dimensions in the image have also been proposed, like the $2D^2PCA$ (Zhou and Zhang, 2005). This last approach is sometimes preferred given its

accuracy at a reduced set of coefficients for image representation. Some other approaches like Gabor filters are used to extract features characterized by spatial frequency, locality and orientation to compensate variations that occur due to change in illumination, pose and expression. Local binary patterns (LBP) is also used to represent the face, and address the problem of lighting variations LBP (Ojala *et al.*, 1996; Marcel and Rodriguez, 2007). Once the feature vector is defined, the most representative subset of features is selected to reduce the dimensionality of the feature space.

The classification or matching module compares the feature vector to facial models within the biometric database, producing a classification score (similarity or distance based). Matching approaches have been divided in *holistic*, *feature-based*, and *hybrid methods* (Zhao *et al.*, 2003). Holistic methods like eigenfaces or Fisherfaces use the whole set of pixels from a face to obtain a smaller representation, and then apply matching for recognition. Feature-based methods like graph matching or hidden Markov models typically use positions and statistics of local features (eyes, nose and mouth). Hybrid methods like modular eigenfaces use local features as well as the whole set of pixels from the face region.

The Spatio-temporal fusion module applies a threshold to the score in order to produce a decision on the input ROIs, that depends on the functionality of the system (e.g. accept/ reject, detect, identify). In that sense, hyperparameters and architecture of the classifier, together with the parameters of the decision module (e.g. user specific threshold), constitute the *biometric model* corresponding to an individual of interest enrolled to the system (biometric database).

Approaches for video based FR combine temporal and feature informations to improve matching performance. Matta and Dugelay categorize existing approaches in those that neglect temporal information, and those that propose strategies to exploit temporal information (Matta and Dugelay, 2009). Two variants can be distinguished among spatio-temporal FR approaches. *Tracking-then-recognition* approaches use segmentation to first crop a detected face, and then track the facial region over time. These approaches typically perform face matching on each frame, and then use majority voting for a final result. *Tracking-and-recognition* approaches

attempt to simultaneously track and recognize, and may combine temporal and spatial information in a unified manner (Barry and Granger, 2007; Ekenel *et al.*, 2010; Zhou *et al.*, 2004), or integrate tracking and recognition within a single algorithm (Franco *et al.*, 2010; Lee *et al.*, 2005; Matta and Dugelay, 2006). This categorization is shown in Table 1.1.

Table 1.1 Categorization of spatio-temporal approaches for FR in video

Temporal Information		**Approach**
Neglected		Eigenfaces (Turk and Pentland, 1991) Fisherfaces (Matta and Dugelay, 2009) Active appearance models (Matta and Dugelay, 2009) Radial basis function neural networks (Matta and Dugelay, 2009) Elastic graph matching (Matta and Dugelay, 2009) Hierarchical discriminative regression trees (Matta and Dugelay, 2009) Unsupervised pairwise clustering techniques (Matta and Dugelay, 2009) Open Set TCM-kNN (Li and Wechsler, 2005) Ensembles of Fuzzy ARTMAP classifiers (Pagano *et al.*, 2012)
Exploited	Tracking-then-recognition	Fisherfaces with facial optical flow (Chen *et al.*, 2001) Dictionary-based face recognition (Chen *et al.*, 2014) Score and quality driven matching (Despiegel *et al.*, 2012) HMM extension for video (Liu and Cheng, 2003)
	Tracking-and-recognition	What-and-Where fusion Neural Network (Barry and Granger, 2007) Local appearance-based face models (Ekenel *et al.*, 2010) Tracking and Recognition using Probabilistic Appearance Manifolds (Lee *et al.*, 2003, 2005) Stochastic tracking and recognition through particle filtering (Zhou *et al.*, 2004) GMMs on unconstrained head motion (Matta and Dugelay, 2006) Recognition confidence and interframe continuity (Franco *et al.*, 2010)

From approaches in literature, it can be seen that recognition performance in video-based approaches is highly degraded by variations in pose, illumination and expression. Spatio-temporal approaches that integrate contextual information over time, and decisions over a sequence of frames in general achieve more robust and accurate performance (Matta and Dugelay, 2009). An interesting case of spatio-temporal combination is the method proposed by Barry et al. (Barry and Granger, 2007), where a what-and-where fusion neural network is used to combine classifier responses (Fuzzy ARTMAP) with location of faces (Kalman filter bank).

1.1.1 Specialized Architectures for FRiVS

Despite the nature of the task, FRiVS has been addressed by only a few authors as an open set problem. This problem consists in managing the fact that there are individuals that might be

rejected, but no information is available on them. Specifically in a surveillance scenario, the amount of unknown (non-target) individuals that may appear in scene usually greatly outnumber the (target) individuals enrolled to the system (See Figure 1.2).

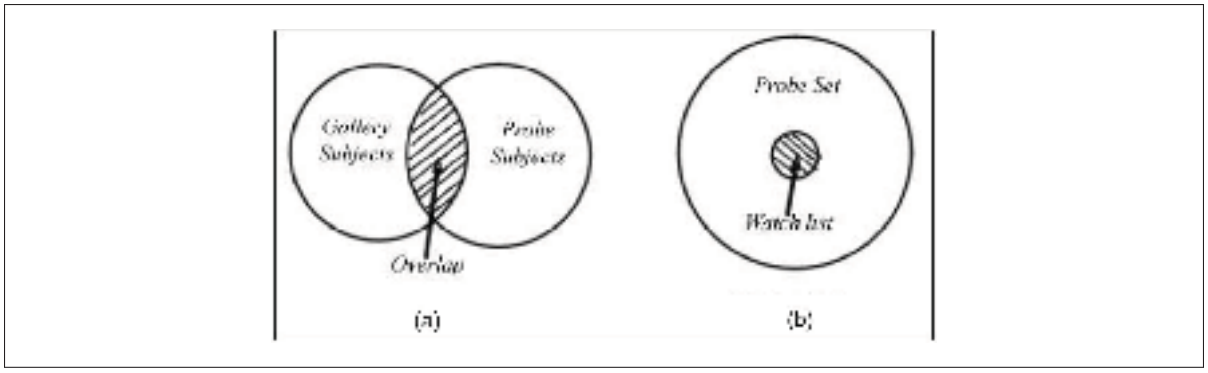


Figure 1.2 Open set (a) and its specialization watch list (b) tasks, extracted from (Li and Wechsler, 2005)

Some performance tests (like FRVT2002) evaluate face recognition algorithms applying a threshold to the output scores of algorithms to decide if the individual is accepted as individual of interest or rejected. Then, identification is performed by comparing such a threshold. Classification architectures that have been found that directly address the open set problem in face recognition are not numerous. For instance, Li and Wechsler use a modified version of *k*-NN called TCM-kNN (Transduction Confidence Machine- k Nearest Neighbors) which considers the new input patterns in order to tune up the rejection threshold (Li and Wechsler, 2005). Tax and Duin propose in (Tax and Duin, 2008) a multi class classifier formed by 1-class binary classifier per class, in which posterior probabilities are normalized to apply a common rejection threshold to all classes, but adapted to each distribution. It is interesting that some results show that the smaller the list of individuals enrolled to the system, the better performance is achieved (Li and Wechsler, 2005; Zhao *et al.*, 2003). This is consistent with idea that individual specific parameters (a sub-system specialized on each individual) might outperform global approaches. This idea is not new, and has been addressed in literature by estimating user specific parameters and thresholds. This strategy leads to a better estimation of facial models

using a specialized matcher for each individual in a modular approach (Bengio and Mariéthoz, 2007; Jain and Ross, 2002; Pagano *et al.*, 2012). And from the classification point of view, it is also known that modular approaches with a classifier per class generally outperform monolithic approaches, specially when data is limited and the classification task is complex (Oh and Suen, 2002; Kapp *et al.*, 2007).

The approach proposed by Kamgar and Parsi, that identifies the decision region(s) in the feature space for each individual face by training a dedicated feed-forward neural network for each individual of interest (Kamgar-Parsi *et al.*, 2011). Another example of such a system is the ensemble of detectors (EoD) designed for each person in a watch list. Non-target samples are retrieved from the CM (database maintained with trajectories from non-target individuals of interest) and the UM (database with training samples from unknown people appearing in scene). Base classifiers are co-jointly trained using a training strategy based on DPSO. It allows for the generation of a diversified pool of ARTMAP neural networks, and trained detectors are then selected and combined using Boolean combination (BC) (Pagano *et al.*, 2012).

In other applications like speaker recognition, the use of the so called "Universal Background Model" to discriminate the target voice from all other sounds is widely used. Also the cohort model uses selected samples from non-target known voices to discriminate known and unknown speakers in open set speaker identification (Brew and Cunningham, 2009, 2010).

From all this, we can see that the surveillance problem can be efficiently addressed as multiple detection problems. This is also consistent with the user-specific parameter optimization, and the use of samples from a negative class (the cohort model) have also provided a better discriminative estimation as it has been applied in speaker verification applications. Besides that, experimental evidence shows that multiple discriminative classifiers generally need less careful calibration and training set selection than generative models (Drummond, 2006). Also ensemble techniques that take advantage of decision level combination are suitable of being used to achieve better performance.

1.1.2 Challenges of FRiVS

Many challenges have been found in FRiVS that remain as a research area. As stated by Zhao et al. in (Zhao *et al.*, 2003), FR from outdoor images of dense scenes, under unconstrained conditions, is still a research problem. This problem has been addressed by considering time information in video-based approaches (Matta and Dugelay, 2009). However noisy sensed data from the complex, changing environment may lead biometric model that does not correspond to the true biometric samples, which affects directly the accuracy of the matching algorithm. Overlapping class distributions due to inter-class similarity also increases the number of false alarms produced by the system. Facial models designed with a limited set of training data from the complex data distribution of faces in feature space are scarcely representative. Even if the facial models are representative, most FR systems assume that face samples in operation are acquired by the same sensor as the used to acquire training data, which is not necessarily true and affect accuracy. Also factors like an inappropriate interaction of the biometric system with the sensor, and inherent scene properties like environmental or temporal changes of the true distribution of faces in feature space, may degrade the accuracy of the system (Rattani, 2010; Poh *et al.*, 2009). The quality of facial models is then a critical issue in the overall biometric application performance. The recognition problem becomes more challenging if we consider that faces do not remain static over time, and present either gradual (e.g. aging) or abrupt (e.g. pose, illumination) changes along the system's operation.

Representative works for FR in video consider the task as an open set problem, where the non-target individuals greatly overcome the target individual of interest. Although FR in video can be addressed with multi-class classifiers, architectures with ensembles of 2-class classifiers (target vs. non-target) take advantage of individual-specific classification parameters, and use the information provided by the abundant non-target samples for increased discrimination.

Another challenge is the retrieval of representative data to design the facial models, which is commonly an expensive (or sometimes not possible possible) activity that require manual labeling of representative images from video archive. Automatic labeling is commonly ad-

dressed in adaptive biometrics by partially-supervised learning techniques, that take advantage of system responses to select label operational samples for system update. The so acquired samples are typically added to a gallery for template matching, but can also be employed for the update of facial models designed with incremental learning classifiers.

The focus of this Thesis lies on the investigation of adaptive FR systems that employ modular architectures with one ensemble of 2-class classifiers per individual of interest enrolled to the system. Adaptive biometric systems may provide initial high performance on face-based video-surveillance, and maintain or increase this performance after adapting with new reference (operational) data. The problem is addressed at the classification stage, considering the accurate design of facial models, and the use of semi-supervised learning to incorporate new knowledge to the biometric database.

1.2 Adaptive Face Recognition

1.2.1 Semi-Supervisd Learning

Many researchers have recently focused on the interesting area of updating biometric models over time employing new acquired data. These adaptive biometric systems can be categorized according to the way class labels are obtained. *Unsupervised approaches* do not require class labels to update biometric models, and a simultaneous recognition and update is performed. On the other hand, *Supervised approaches* use only labeled data previously acquired in an off-line update. Approaches in which biometric models are built supervised, and unsupervised adaptation is performed online, are also called *partially-supervised* or *semi-supervised*.

Table 1.2 shows different approaches to adapt facial models as new data becomes available, either from daily operations or security reports.

It is important to note that even if matching algorithm is a supervised classifier, the construction or adaptation of biometric models can be performed in an unsupervised way. This is the case of the approach described in (Mou *et al.*, 2006; Mou, 2010), where author use the classification

Table 1.2 Different adaptive approaches for FR and their methods to build and adapt facial models

Approach	Build BM	Adaptation BM	Matcher
Eigenfaces (Turk and Pentland, 1991)	Supervised	Unsupervised	1-NN
(Okada *et al.*, 2001)	Supervised	Unsupervised	Elastic Bunch Graph
(Mou *et al.*, 2006; Mou, 2010)	Unsupervised	Unsupervised	Distance based (Black box: FaceVACS)
(Rattani *et al.*, 2008b; Rattani, 2010)	Supervised	Unsupervised	Graph Matching
(Singh *et al.*, 2010)	Supervised	Unsupervised	2ν-Online Granular SVM
(Connolly *et al.*, 2010a)	Supervised	Supervised	Fuzzy ARTMAP
(Connolly *et al.*, 2010b)	Supervised	Supervised	Ensemble of Fuzzy ARTMAP
(Franco *et al.*, 2010)	Supervised	Unsupervised	Distance-based template matching
(Ekenel *et al.*, 2010)	Supervised	Unsupervised (adapt thresholds)	*k*NN (DTM and DT2ND)
(De-la Torre *et al.*, 2014a)	Supervised	Unsupervised	Ensembles of PFAM classifiers

algorithm as a black box that produces a score (distance based), and the decision of incorporating or not a new sample is based on rules that compare a threshold to its corresponding score. In a human centered scenario, where new labeled data from individuals of interest becomes available (e.g. due to security reports), semi-supervised approaches seem more interesting since human knowledge can be combined with human expertise.

1.2.2 Adaptive Biometrics

In the literature, several approaches allow for supervised adaptation providing reliable results (De-la Torre *et al.*, 2012a; Connolly *et al.*, 2012; Tax and Duin, 2008), and yet obtaining labeled reference samples is costly or impractical. To overcome this difficulty, some *semi-supervised* methods have been introduced for automatic template updates (Roli and Marcialis, 2006; Franco *et al.*, 2010; Roli *et al.*, 2007, 2008; Okada *et al.*, 2001; Rattani *et al.*, 2008a, 2009b). This chapter focuses on the semi-supervised updating of biometric models. *Self-training* and *co-updating* are two well-known algorithms for semi-supervised adaptation using template matching.

In *self-update* methods (Roli *et al.*, 2007), the biometric models are first designed storing samples from a labeled data set D_L in a template gallery $\mathscr{G}$. Prediction is possible by applying a decision threshold γ^d to the similarity score produced after template matching. Then, during

operations, similarity scores are produced for the unlabeled samples, and those with a high degree of confidence (surpassing an updating threshold $\gamma^u \geq \gamma^d$), are integrated to the gallery $\mathscr{G}$, thereby updating the corresponding biometric models. The notion of "high degree of confidence" is subjective, and depends on both the matching algorithm and the application domain, but an update threshold higher or equal than the prediction threshold is commonly used. This procedure is detailed in Algorithm 1.1.

Algorithm 1.1: Self-update algorithm to adapt a gallery for template matching

Input :
  $\mathscr{G} = \{t_1, ..., t_N\}$ // Gallery with initial templates
  $D = \{d_1, ..., d_L\}$ // Unlabeled adaptation set
Output :
  $\mathscr{G}' = \{t_1, ..., t_N, ..., t_M\}, M \geq N$ // Updated template gallery
Estimate threshold $\gamma^u \geq \gamma^d$ *for the templates in* $\mathscr{G}$
$\mathscr{G}' \Leftarrow \mathscr{G}$ // Initialize with $\mathscr{G}$
// For all samples $d_l \in D$
for $l = 1, ..., L$ **do**
  // For all templates in the gallery $t_l \in \mathscr{G}$
  for $n = 1, ..., N$ **do**
    $s_{n,l} \Leftarrow similarity\_measure(d_l, t_n)$ // Compute score against all samples in $\mathscr{G}$
  $s_l \Leftarrow \max\{s_{n,l} : n = 1, ..., N\}$
  if $s_l > \gamma^u$ **then**
    $\mathscr{G}' \Leftarrow \mathscr{G}' \cup d_l$ // Include the sample surpassing γ^u in the new data set

Co-update is a semi-supervised learning strategy adapted for use with two diversified matchers with galleries specialized on distinct biometric traits, which are designed to improve performance mutually (Roli *et al.*, 2007). For example, in (Roli *et al.*, 2007), authors propose the use of fingerprints and the face, using co-training for semi-supervised updates of the facial and fingerprint models. Algorithm 1.2 presents the co-training algorithm. The procedure starts with the design of the two matchers with the labeled templates in galleries $\mathscr{G}_1$ and $\mathscr{G}_2$, and selecting ad-hoc the thresholds for decision (γ_1^d and γ_2^d) and update (γ_1^u and γ_2^u). Once the unlabeled sets D_1 and D_2 are collected, both matchers are used to label the samples, and those with high degrees of confidence (at least in one of the matchers) are added to the updated galleries $\mathscr{G}_1'$ and $\mathscr{G}_2'$. Also the decision and update thresholds are be updated over time in accordance with the newly acquired data. A potential advantage of the co-update algorithm is that it can retrieve

update samples that are not typical of the distribution of target data from a single trait, allowing adaptation to diverse, possibly abrupt changes.

The advantages of adapting a biometric system using operational data carries an inherent risk. There exists a trade-off between the false updates and false rejections that affect of performance. A conservative threshold (or other parameters in the biometric model) may allow a system without false updates, but also a system that is never adapted to changes in the environment. Conversely, a less conservative threshold may contribute to increase in the number of false updates and the inherent deterioration of biometric models. Following this reasoning, we can easily see that a good selection of adaptation criteria (decision threshold) is crucial in the design of the system.

Algorithm 1.2: Co-update algorithm to adapt a gallery for template matching

Input :
$\mathscr{G}_1 = \{t_1^1, ..., t_{N_1}^1\}$
and $\mathscr{G}_2 = \{t_1^2, ..., t_{N_2}^2\}$ // Galleries with initial templates
$D_1 = \{d_{1,1}, ..., d_{L,1}\}$
and $D_2 = \{d_{1,2}, ..., d_{L,2}\}$ // Unlabeled adaptation sets, $d_{l,1}$ corresponds to $d_{l,2}$
Output :
$\mathscr{G}_1' = \{t_1^1, ..., t_{N_1}^1, ..., t_{M_1}^1\}$,
$M_1 \geq N_1$ // Updated galleries for both modalities
$\mathscr{G}_2' = \{t_1^2, ..., t_{N_2}^2, ..., t_{M_2}^2\}$, $M_2 \geq N_2$
Estimate thresholds $\gamma_1^u \geq \gamma_1^d$ *and* $\gamma_2^u \geq \gamma_2^d$ *for the* $\mathscr{G}_1$ *and* $\mathscr{G}_2$ *respectively*
// For each gallery G_i, $i = 1,2$
for $i = 1,2$ **do**
  $\mathscr{G}_i' \Leftarrow \mathscr{G}_i$ // Initialize with templates in the gallery i
  // For all samples $d_{l,i} \in D_i$
  for $l = 1, ..., L$ **do**
    // For all templates in the gallery $t_{n,i} \in \mathscr{G}_i$
    for $t_{n,i} \in \mathscr{G}_i$, $n = 1, ..., N_i$ **do**
      $s_{n,l,i} \Leftarrow similarity\_measure(d_{l,i}, t_{n,i})$ // Compute score for all $d_n \in D_i$
    $s_{l,i} \Leftarrow \max\{s_{n,l,i} : n = 1, ..., N_i\}$
    if $s_{l,i} > \gamma_i^u$ **then**
      $j \Leftarrow \mod(i+1,2)+1$ // Samples added to the complementary gallery
      $\mathscr{G}_j' \Leftarrow \mathscr{G}_j' \cup d_{l,j}$

Other semi-supervised approaches take advantage of neural or statistical classifiers in the construction of biometric models. For instance, in (Okada *et al.*, 2001), a view representation that combines facial and torso-color histograms was used with bunch graph matching for adaptive person recognition. The system is capable of updating existing biometric models and to automatically enroll unknown individuals based on a double thresholding strategy. Update was performed on operational video streams that provide high sequence-to-entry similarity, measure of confidence. The sequence-to-entry similarity is the average of maximum frame-to-entry similarity values, which in turn was defined as the maximum similarity value over all facial representations in a database entry (Okada *et al.*, 2001). Bayesian networks were also used to recognize facial expression and detect faces using a stochastic structure search algorithm (Cohen *et al.*, 2004). This approach combined labeled and unlabeled samples to train the Bayesian networks, and seek for the Bayesian network structure that provided the minimum probability of error, using maximum likelihood estimation. SVMs with locality preserving projections have also been combined to update facial models, by incorporating information from operational ROIs taken from video (Lu *et al.*, 2010). The algorithm first builds a data model of a video sequence, and then uses semi-supervised locality preserving projections to assemble a graph with the geometrical structure of the feature space of faces.

MCSs have also been used in conjunction with the co-training and self-training. In (Didaci and Roli, 2006), for instance, an ensemble of five classifiers was trained with two different diversity generation techniques (bootstrap and the training of different classifiers). These techniques are based on a re-training schema for biometric model updates, and improve accuracy by 18% using the product rule for combination. Another modification of the co-training algorithm for MCS was proposed for updating only unlabeled samples that produced high confidence (El Gayar *et al.*, 2006). The five patterns with highest probability of belonging to the specific person, were selected as the most confident. This system was tested with 3 non-homogeneous classifiers in the ensemble, and provided the highest performance with a voting combination scheme. Finally, a semi-supervised classification schema based on random subspace dimensionality reduction was proposed for graph-based semi-supervised learning. In this approach,

a kNN graph is built in each processed random subspace, and semi-supervised classifiers are trained on the resulting graphs, using majority voting rule for combination (Yu *et al.*, 2012).

1.2.3 Challenges of Adaptive FR Systems

Adaptive FR systems that perform partially-supervised learning commonly employ the classifier predictions to assign unlabeled data to a determined class, and add those recently labeled patterns to a gallery for template matching. Adaptive biometric systems for partially-supervised learning provide two main strategies. Self-update based on the output scores of a single matcher, whereas co-update employs two diversified matchers with galleries specialized on distinct biometric traits, which are designed to improve performance mutually.

In this Thesis, co-update and self-update strategies are explored for FRiVS, and a strategy is proposed for self-update of facial models based on face trajectories. Several challenges emerge, including how to combine the tracking and classification information, how to correctly select the optimal decision and update thresholds, and how the new samples are learned to update the facial model while avoiding the knowledge corruption.

1.3 Incremental and On-Line Learning of Classifiers

Adaptation of biometric models for FR has been addressed by adapting the feature space (amount and orientation of feature vectors that generate the feature space), or the classifier (internal knowledge of the classification algorithm) (Ozawa *et al.*, 2005). In this Thesis, the adaptation is addressed using ensemble-based learning and a *learn-and-combine* strategy that allows to integrate information of changes in feature space, avoiding the corruption of knowledge of the facial models.

The design of neural or statistical classifiers involves the estimation of a mapping $f : \mathbf{a} \rightarrow \Omega$, between the feature vector $\mathbf{a} \in \mathbb{R}^I$ and class labels $\Omega = \{C_1, C_2, ..., C_K\}$. Generative classifiers tend to approximate probability distribution functions of the different classes from training data (e.g. k-NN, RBF). On the other hand, discriminative classifiers approximate classification

boundaries by finding discriminant functions (MLP, SVM). In discussion of which performs better no one can be favored for a general case. However it has been shown that generative classifiers commonly require more training data and require more careful calibration to overcome multiple discriminative classifiers (Drummond, 2006). Some algorithms like the ARTMAP family provide characteristics from both approaches. These neural networks are characterized by fast supervised, unsupervised and incremental learning from limited amounts of data, and comparable classification accuracy when compared to state of the art classification algorithms (Lerner and Guterman, 2008).

In real world applications, it is typical that new data becomes available over time, and the knowledge of a pattern recognition systems may be adapted to maintain accuracy. Figure 1.3 shows the scenario where a classifier performs supervised learning of data incrementally. This capacity of learning data incrementally allows classifiers to update their internal knowledge.

According to Polikar et al. (Polikar *et al.*, 2001), an incremental learning algorithm should be able to learn additional knowledge from a new block of data without requiring storage or access to previously learned data, should preserve previously acquired knowledge, and should be able to learn new classes that may be introduced with new data. In Figure 1.3 a classifier hyp_1 is initially designed using a set of user-defined *hyperparameters* and a limited amount of training data D_0. Different datasets $D_1, ..., D_t$ from existing and/or new classes may become available over discrete instants of time $t = 1, 2, ...$, and parameters of hyp_1 are updated to hyp_2 by incorporating data samples from D_1. Similarly hyp_2 is updated to hyp_3 on the basis of data D_2 and so on.

In literature, three types of pattern classification algorithms for IL have been proposed (see Table 1.3). The first consists on classifiers that have been designed with the inherent ability to perform supervised incremental learning, and examples this category are the ARTMAP (Lerner and Guterman, 2008) and Growing Self-Organizing (Fritzke, 1996) families of neural networks. The second category is composed of some well-known pattern classifiers, such as Support Vector Machines (SVM), the Multi Layer Perceptron (MLP) and the Radial Basis

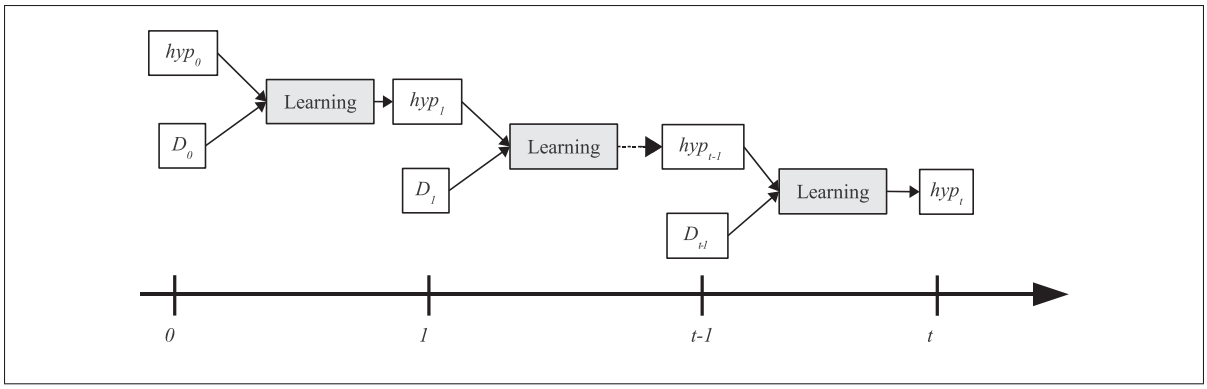


Figure 1.3 Incremental learning (IL) scenario, where new data D_t is learned by the classifier to update its parameters and architecture, extracted from (Connolly *et al.*, 2008)

Function (RBF) Neural Network, which have been adapted to perform supervised incremental learning. Finally the third consists on ensembles of classifiers that may update parameters or architecture when new data becomes available.

Table 1.3 Classifiers that are capable of Incremental learning

IL	Approaches
INHERENT	**ARTMAP based**, categorized according to category class activation: *Winner-Take-All.* ARTMAP(Carpenter *et al.*, 1991), Fuzzy ARTMAP (Carpenter *et al.*, 1992), ARTMAP-IC(Carpenter and Markuzon, 1998). *Distributed.* ART-EMAP(Carpenter and Ross, 1995), distributed ARTMAP (Carpenter and Milenova, 1999). *Probabilistic.* PFAM (Lim and Harrison, 1995, 1997), PSFAM (Jervis *et al.*, 1999), Gaussian ARTMAP (Williamson, 1996), hypersphere ARTMAP (Anagnostopoulos and Georgiopulos, 2000), Ellipsoid ARTMAP (Anagnostopoulos and Georgiopoulos, 2001), boosted ARTMAP (Verzi *et al.*, 1998), μARTMAP (Gomez-Sanchez *et al.*, 2002) and Bayesian ARTMAP (Vigdor and B., 2007). *Hybrids.* Default ARTMAP 2 (Amis and Carpenter, 2007) (WTA training and distributed activation).
	Growing Self-Organizing Feature Maps: Growing Neural Gas (Fritzke, 1996) and Hybrid Self-Organizing Neural Gas (Graham and Starzyk, 2008)
MODIFIED	**Support Vector Machines.** Incremental SVM and variants (Syed *et al.*, 1999; Ruping, 2001; Li and Huang, 2002); (Diehl and Cauwenberghs, 2003), 2ν-Online Granular Soft SVM (Singh *et al.*, 2010)
	Multi Layer Perceptron. Incremental back-propagation algorithm (Fu *et al.*, 1996; Wang and Kuh, 1992); (Osorio and Amy, 1999; Lee, 1990).
	Radial Basis Function neural networks. Resource allocation network (RAN) (Platt, 1991), Kalman filter (RANEKF) (Kadirkamanathan and Niranjan, 1993), MRAN(Yingwei *et al.*, 1997), Growing and Pruning-RBF (Salmeron *et al.*, 1999, 2001)
ENSEMBLE	**Learn++.** Learn++ (Polikar *et al.*, 2002), Learn++.NSE, Learn++.MT (Muhlbaier *et al.*, 2004), Learn++NC (Muhlbaier *et al.*, 2009), Learn++.UDNC (Ditzler *et al.*, 2010), Learn++.MF (Polikar *et al.*, 2010)
	Adaptive Fuzzy ARTMAP ensemble. Connolly et al (Connolly *et al.*, 2010b)

Classifiers designed with the inherent ability to perform supervised incremental learning are inspired on the well-known self-organizing neural networks (SONNs). The unsupervised learning paradigm used in SONNs is related to clustering, since it permits the assignment of adaptively defined categories to unlabeled patterns. Modifications of typical classifiers including

SVM or MLP have been also proposed to give them the capacity of learning new data after classifiers are in operation. Ensemble based approaches in general outperform single monolithic classifier, which can be explained with the degradation that may occur while updating the state of knowledge of classification algorithms. This state of knowledge may conduct the classifier to be trapped in local optima.

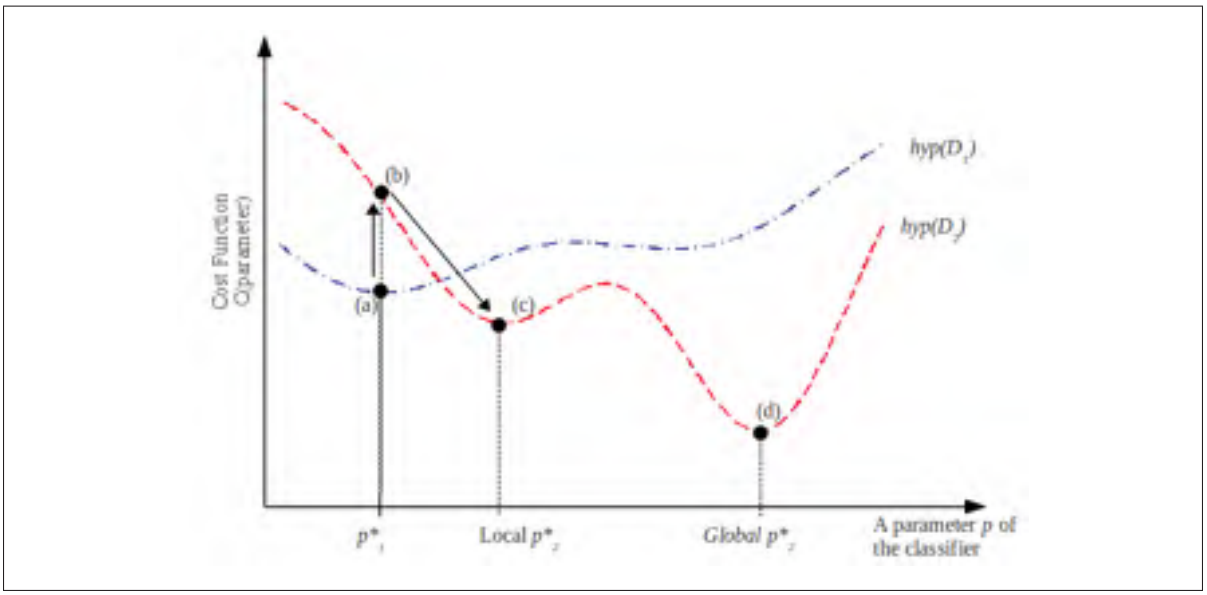


Figure 1.4 Illustration of knowledge corruption with monolithic approaches for incremental learning
extracted from (Khreich *et al.*, 2012)

In Figure 1.4, a monolithic classifier is initially trained on block D_1 sampled from a probability distribution, producing the costs represented by hyp_1 on a determined parameter p of the classification algorithm. Given that the first time of training is equivalent to batch learning, it is possible to reach the global minimum (a) that fits with samples in D_1. Then hyp_1 is updated on the new data block D_2, and the new minimum (c) is estimated instead of (d), because of the current parameters of the classifiers are affected by the previous estimation (b). This phenomenon is called knowledge corruption, and ensemble based classifiers avoid this problem by maintaining several solutions to the classification problem. This allows this techniques to be less likely to fall in local minimums. In fact, performing incremental learning with ensemble

based techniques, using a learn and combine approach has been successfully applied in some areas (Polikar *et al.*, 2001; Polikar, 2006; Arandjelovic and Cipolla, 2006).

This Thesis focus on ensemble-based techniques for designing the classification system for incremental learning scenarios. The following sub-sections present detailed descriptions of the techniques employed for classification, including the ARTMAP classifier, Boolean combination and the DPSO training strategy for generation of classifiers.

1.3.1 Fuzzy ARTMAP

The fuzzy ARTMAP neural network is a member of the ARTMAP family, which integrates a fuzzy ART module to process both analog and binary-valued input patterns to the original ARTMAP architecture (Carpenter *et al.*, 1992). Simplified architecture of the Fuzzy ARTMAP classifier is shown in Figure 1.5. Two fully connected layers of nodes (F_1 and F_2) constitute the main ART network, and a third layer (F^{ab}) is used for training by using back propagation strategies. Connections between layers are associated with different weights $\mathbf{W}$ and $\mathbf{W}^{ab}$. Tuning parameters of this classifier are the learning rate β, choice α, match tracking ε and the baseline vigilance parameter $\bar{\rho}$.

Real valued weights $W = \{w_{ij} \in [0,1] : i = 1,2,...,M; j = 1,2,...,N\}$ are associated to the connections between the input layer F_1 and the competitive layer F_2. Each node j from F_2 represents a recognition category that learns a prototype vector $\mathbf{w_j} = (w_{1j}, w_{2j}, ..., w_{Mj})$. Learned connections between nodes from layer F_2 to F^{ab} are associated with binary weights $\mathbf{W}^{ab} = \{w^{ab}_{jk} \in [0,1] : j = 1,2,...,N; k = 1,2,...,L\}$. The link that joins the F_2 j node with one of the L output classes, is the vector $\mathbf{w}^{ab}_j = (w^{ab}_{j1}, w^{ab}_{j2}, ..., w^{ab}_{jL})$. Input patterns for batch supervised training mode are pairs $(\mathbf{a}, \mathbf{t})$ where $\mathbf{a}$ is the pattern itself and $\mathbf{t}$ is its binary supervision pattern set.

Activation of each node j in layer F_2 from layer F_1 is determined by the *Weber Law choice function* given by

$$T_j(\mathbf{A}) = \frac{|A \wedge \mathbf{w}_j|}{\alpha + |\mathbf{w}_j|}, \tag{1.1}$$

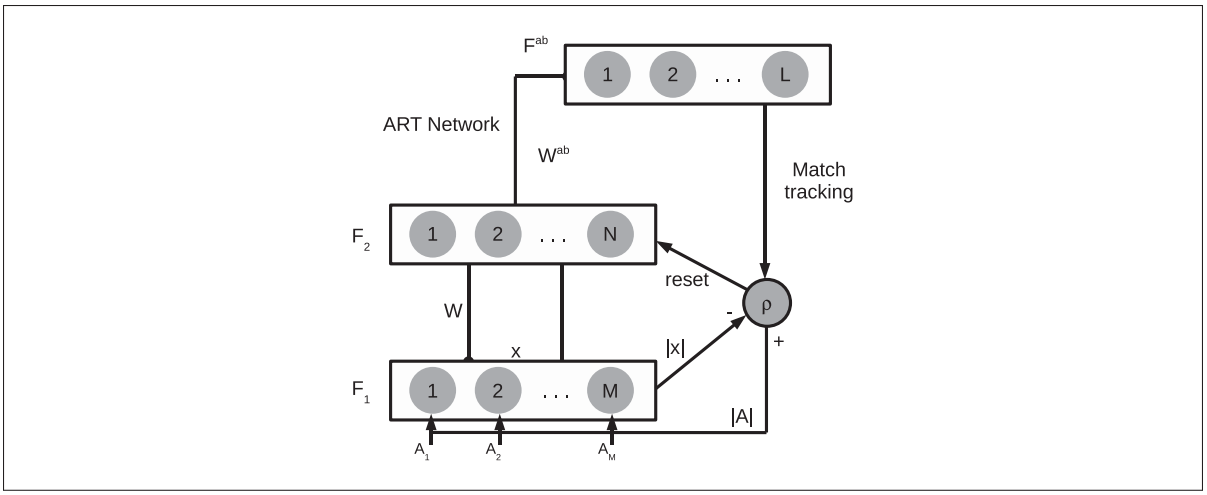


Figure 1.5 Simplified architecture of ARTMAP Neural Networks
extracted from (Granger *et al.*, 2007)

where $|\ |$ is defined by $|\mathbf{w}_j| = \sum |w_{ij}|_{i=1}^{M}$, $\wedge$ is the fuzzy AND operator, $(\mathbf{A} \wedge \mathbf{w}_j) = min(A_i, w_{ij})$.

After F_2 nodes are activated, the layer produces a binary pattern of activity $\mathbf{y} = (y_1, y_2, ..., y_N)$ by applying the *winner-take-all* strategy, where only the node $j = J$ with the greatest activation value $J = argmax\{T_j : j = 1, 2, ..., N\}$ remains active. Then, the degree of match between expectation vector $\mathbf{w}_J$ and the input vector $\mathbf{A}$ are compared against the vigilance parameter with the *vigilance test* (1.2).

$$\frac{|\mathbf{A} \wedge \mathbf{w}_J|}{|\mathbf{A}|} = \frac{|\mathbf{A} \wedge \mathbf{w}_J|}{M} \geq \rho. \tag{1.2}$$

Depending on this test it is said that resonance occur, or the network inhibits the active F_2 node and searches for another node that passes the test. Once this pattern $\mathbf{w}_J$ is found, F^{ab} layer produces a binary pattern of activity $\mathbf{y}^{ab} = \mathbf{t} \wedge \mathbf{w}_J^{ab}$, and the most active F^{ab} node ($K = k(J)$) which constitutes the prediction class. Prediction function uses the *Winner-Take-All* strategy, obtaining a competing score with (1.3).

$$S_k^{ab}(\mathbf{y}) = \sum_{j=1}^{N} y_j w_{jk}^{ab} \tag{1.3}$$

When a wrong prediction is obtained compared to the target class **t** of the pattern, the vigilance parameter ρ must be updated by

$$\rho = \frac{|\mathbf{A} \wedge \mathbf{w}_J|}{M} + \varepsilon \tag{1.4}$$

Fuzzy ARTMAP learning algorithm can be summarized in the following five steps:

1. *Initialization.* F2 nodes are uncommitted[1], w_{ij} weights are initialized to 1, w_{ij}^{ab} weights are initialized to 0, and parameters $\alpha > 0$, $\beta \in [0,1]$, $0 < \varepsilon << 1$ and $\bar{\rho} \in [0,1]$ are set.

2. *Input Pattern Coding.* For each presented input training pattern $(\mathbf{a},\mathbf{t}) = (a_1, a_2, ..., a_m, t_1, t_2, ..., t_L)$, the complement coding of **a** is evaluated. F_1 input is then transformed to $\mathbf{A} = (\mathbf{a}, \mathbf{a}^c) = (a_1, a_2, ..., a_m; a_1^c, a_2^c, ..., a_m^c)$, where $a_i^c = (1 - a_i)$, a_i must be normalized ($a_i \in [0,1]$). Parameter ρ is set to $\bar{\rho}$.

3. *Prototype Selection.* Pattern **A** activates F_1 and is propagated through **W** to F_2. Activation of each j-node in F_2 is determined by (1.1). Winner node $J = argmax\{T_j : j = 1,2,...,N\}$ propagates its top-down expectation and vigilance test is performed with (1.2). If test is passed, J remains active and resonance is said to occur. Otherwise the active F_2 node is inhibited and the net searches another J node that passes the test. If such a node does not exist, a new F_2 node is committed and the net goes to learning (step 5).

4. *Class Prediction.* The F^{ab} layer generates $\mathbf{y}^{ab} = (y_1^{ab}, y_2^{ab}, ..., y_L^{ab}) = \mathbf{t} \wedge \mathbf{w}_J^{ab}$ and gets the class prediction $K = argmax\{y_k^{ab} : k = 1,2,...,L\} = k(J)$. The score function (1.3) is evaluated and,if node K constitutes a wrong prediction, ρ is updated with (1.4). This search continues until either an uncommitted F_2 node becomes active or a node J that previously learned the correct class prediction K becomes active.

5. *Learning.* When a pattern **a** produces resonance with an F_2 committed node J, or an uncommitted node becomes active, prototype vector $\mathbf{w}_J$ is updated according to

$$\mathbf{w}_j' = \beta(\mathbf{A} \wedge \mathbf{w}_J) + (1 - \beta)\mathbf{w}_j. \tag{1.5}$$

[1] An F_2 node becomes committed when is selected to code an input vector **a**, and then linked to an Fab node.

1.3.2 PFAM Neural Classifier

Of special interest are probabilistic variants of the Fuzzy ARTMAP algorithm like the proposed by Lim and Harrison, which uses the discriminative learning strategy of Fuzzy ARTMAP, and the generative prediction of Probabilistic Neural Networks (PNN)(Lim and Harrison, 1995, 1997). The list below summarizes the few differences in the learning and prediction phases of both approaches.

1. *Learning Phase*. A FAM structure is used as supervised, clustering algorithm: F_2^a nodes are used to code prototype patterns and update centers of mass ($\mathbf{w}_j^{a-c} = (w_{1j}^{a-c}, w_{2j}^{a-c} \ldots w_{Mj}^{a-c})$). It also encodes the frequency counts associated with F_2 node activations by using the frequency counts in W^{ab}.

2. *Prediction Phase*. The PNN is used to perform probability estimation, and Bayes' decision theorem is applied to select the class whith maximum a posteriori probability, or use another risk-weighted classification rule. In this way, each category j is represented as a hyper-spherical Gaussian pdf according to

$$g_j(\mathbf{a}) = \frac{1}{(2\pi)^{M/2}\sigma_j^M} \exp\Big(-\frac{(\mathbf{a}-\mathbf{w}_j^{a-c})^T(\mathbf{a}-\mathbf{w}_j^{a-c})}{2\sigma_j^2}\Big), \tag{1.6}$$

   where the variance σ_j is the ratio of the sqared minimum Euclidean distance between $\mathbf{w}_j^{a-c}$ and any other center vector, to the value of an overlap parameter $r > 0$.

1.4 Adaptive Ensembles

Ensemble-learning techniques combine classifiers with diversity of opinions to increase classification performance. The design process can be divided into three main steps – generation of a pool of base classifiers, selection and fusion of classifiers (Duda *et al.*, 2001; Kuncheva, 2004; Zenobi and Cunningham, 2001; Britto *et al.*, 2014). The first step allows to train base classifiers with diversity of opinions, and the last two take advantage of this diversity to produce more

accurate predictions. Diversity can be created by employing distinct classifiers, train distinct instances of a classifier with different initial conditions (parameters), or using different training sets (Kuncheva, 2004).

Representative examples of ensemble methods are bagging, boosting, random subspaces, which employs different training sets of data or features from the training set to build distinct base classifiers (Kuncheva, 2004; Kittler, 1998). An example of diversity generation by various parameters is the work of Connolly et al. (Connolly *et al.*, 2012), which takes advantage of diversity in the hyperparameter space of classifiers to produce useful diversity of opinions. Examples of selection strategies are greedy search, clustering-based methods and ranking-based methods, and examples of fusion strategies can be divided in feature-based, score-based and decision-based (Tao and Veldhuis, 2008).

1.4.1 Generation of Pools

There are different ways to generate a diverse ensemble. The use of different training datasets for different classifiers usually take advantage of resampling techniques with (bootstrapping or bagging) or without (jackknife or k-fold data split) resampling. Using different internal parameters for different classifiers, or even different algorithms is also a common strategy to produce disagreement between ensemble members. The use of different features to train each ensemble member is also used and referred as *random subspace method*. Some measures of diversity used in literature include diversity, correlation, Q-Statistic, disagreement and double fault measures, entropy, Kohavi-Wolpert variance and difficulty (Polikar, 2006). An interesting technique to generate (and maintain) diversity in the optimization space takes advantage of PSO techniques and produces an heterogeneous ensemble of ARTMAP classifiers. Such a technique has been successfully applied in face recognition applications (Connolly *et al.*, 2010b).

Of special interest are the PSO-based training strategies that co-jointly optimize parameters and architecture of 2-class binary classifiers, according to both accuracy and resources. First introduced by Kennedy and Eberhart in 1995, Particle Swarm Optimization (PSO) is a population-

based stochastic optimization technique that takes advantage of Artificial-Life ideas like bird flocking or fish schooling, together with evolutionary computation to search for the global best of nonlinear functions. Each particle corresponds to a single solution in the optimization space, and the population of particles is called a swarm. In this strategy, the best position and the best position of its surrounding corresponding to each particle are kept in memory.

The main idea can be formally stated as follows. Diversity is maintained employing a local neighborhood topology and by dynamically creating subswarms around masters (particles that are their own best position amongst their neighborhood). Particles that are not part of any sub swarms are called free particles and are allowed to move by themselves. The position of the particles that are members of a subswarm are updated with

$$\begin{aligned} h_n(t+1) &= h_n(t) + w_0(h_n(t) - h_n(t-1)) \\ &\quad + r_1 w_1/2(h^*_{master} - h_n(t)) \\ &\quad + r_2 w_1/2(h^*_n - h_n(t)) \end{aligned} \tag{1.7}$$

where $h_n(t+1)$ is the position of particle n in the optimization space at iteration (t+1), w_0 and w_1 are inertia weights, r_1 and r_2 are random numbers generated at each iteration, $h_n(t)$ and h^*_n are respectively the current position of the subswarm master's personal best (social influence) and particle n personal best (cognitive influence). Free particles move only according to their own cognitive influence using:

$$\begin{aligned} h_n(t+1) &= h_n(t) + w_0(h_n(t) - h_n(t-1)) \\ &\quad + r_3 w_1(h^*_n - h(t)), \end{aligned} \tag{1.8}$$

where r_3 is another random number generated at each iteration. The global best particle is referred to as *gbest*, and in case there is a tie for the global best position, the particle with the smallest index wins.

The dynamic PSO (DPSO) training algoritm employed to generate base classifiers was proposed by Connolly et al. (Connolly *et al.*, 2010b), and is based on the DNPSO algorithm developed in (Nickabadi *et al.*, 2008). Algorithm 1.3 describes the DPSO-based incremental learning strategy for co-optimization of hyperparameters, weight and architecture of the fuzzy ARTMAP neural network. Given new learning data block Dt , it produces the optimal set of hyperparameters and network using a particle swarm with N particles, and N + 2 fuzzy ARTMAP neural networks with one network per particle $PFAM_n$, used to preserve the model associated to the best position of that particle ($\mathbf{h}_n^*$), one temporary neural network used for the fitness estimation during the algorithm ($PFAM_{estimation}$), and one optimal network ($PFAM_{optimal}$).

1.4.2 Selection and Fusion

From the point of view of information fusion, the fusion of classifier outputs can be achieved at matching score and decision levels (See Figure 1.6).

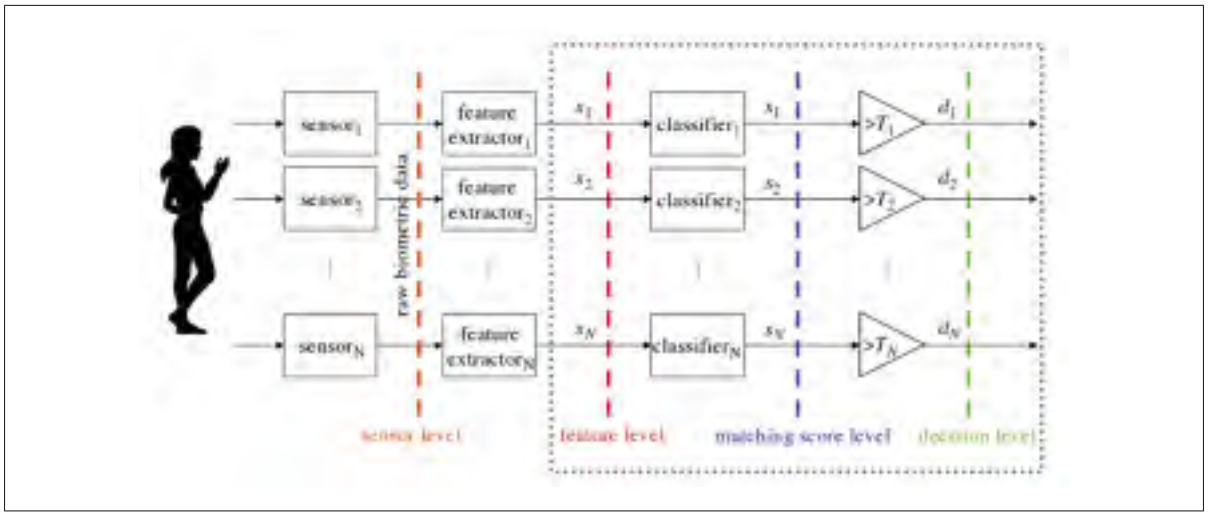


Figure 1.6 Different information fusion levels in biometric systems extracted from (Tao and Veldhuis, 2009)

The **Matching score fusion level** is probably the most used and studied. Three categories are basically studied in literature, including *transformation-based*, *density-based* and *classifier*

Algorithm 1.3: DPSO learning strategy to generate diverse classifiers (Connolly *et al.*, 2010b)

Input : A set of particles (swarm with DNPSO parameters), and neural networks: $PFAM_n$, where $1 \leq n \leq N$, $PFAM_{estimation}$, and $PFAM_{optimal}$, and data D_t for learning.

Output : $PFAM_{optimal}$ (Weights and architecture obtained with the optimal h) and $PFAM_n$ where $1 \leq n \leq N$ (Set of PFAM neural networks associated to the best position of each particles).

Initialization:

1: Set the swarm parameters (N, w_0, w_1).
2: Randomly initialize particles positions for $t = 0$ and $t = -1$ within their range.
3: Initialize $PFAM_{optimal}$ and all $PFAM_n$, where $1 \leq n \leq N$.
4: Set PSO iteration counter at t = 0.

Upon reception of a new data block D_t , the following incremental process is initiated:

Update the fitness of networks associated to the personal best positions:

5: **for** each particle n, where $1 \leq n \leq N$ **do**
6: $PFAM_n \Leftarrow PFAM_{optimal}$
7: Training of $PFAM_n$ with validation using D_t and D_t^v, and $f(h_n^*, t)$ estimation using D_t^f.

Optimization process:

8: **while** DNPSO did not reach stopping condition **do**
9: Define the subswarms and update position of each particle with Eq. 1.7 and 1.8.
10: **for** each particle n, where $1 \leq n \leq N$ **do**
11: $PFAM_{estimation} \Leftarrow PFAM_{optimal}$
12: Training of $PFAM_{estimation}$ with validation using D_t and D_t^v, and
13: $f(h_n(t), t)$ estimation using D_t^f.
14: **if** $f(h_n(t), t) > f(h_n^*, t)$ **then**
15: $h_n^* \Leftarrow h_n(t)$
16: $f(h_n^*, t) \Leftarrow f(h_n(t), t)$
17: $PFAM_n \Leftarrow PFAM_{estimation}$
18: $t = t + 1$

Define the neural network with the highest accuracy:

19: $PFAM_{optimal} \Leftarrow PFAM_{gbest}$

based. In *transformation-based* fusion, all component matching scores are first transformed or normalized, and then simple scalar functions are applied to produce a new matching score. Some commonly used functions are product, sum, mean, max, etc. And under some ideal situations, they can achieve statistically optimal performance in the Neyman-Pearson sense. *Density-based* schemes are based in the estimation of joint densities of matching scores. Fusion is done by using statistical tests like likelihood ratio. Optimal performance of these schemes could be achieved when a large number of representative training matching scores are available. *Classifier-based* fusion schemes concatenate the matching scores as a new feature vector,

and an additional classifier is trained on this representation. However it has been concluded in a study by Roli et al. (Roli *et al.*, 2002), that trained rules could not provide significant advantages over fixed rules, specially in ensembles with members that achieve different performances.

Decision level fusion techniques include the well known majority voting, weighted majority voting, Bayesian decision fusion and the Dempster-Shafer theory of evidence. Operating points in this space can be selected according to their performance characterized by the pair (fpr, tpr). When only a non-continuous curve is available, it can be always possible to interpolate between two points. Scott et al (Scott *et al.*, 1998) use the so called Maximum Realizable ROC (or Convex Hull), interpolating between non-existing points. Also the selection of specific thresholds of different classifiers to pick specific operating points in ROC space, to be combined with AND and OR rules has been studied by Haker et al. in (Haker *et al.*, 2005). Barreno et al. (Barreno *et al.*, 2008) use AND and OR rules to combine all points in ROC space produced by classifiers, and find an optimal ROC curve given by such a combination of decision rules. This way of combining binary classifiers is optimal in the Neyman-Pearson sense, however a high complexity in their algorithm (2^{2^n} possible Boolean rules). Tao and Veldhuis (Tao and Veldhuis, 2009) propose a decision level fusion scheme in which thresholds at score level are tunned to fix a false-acceptance rate, such that the decision false rejection rate is minimal. This threshold optimized scheme takes advantage of AND and OR rules, and its main disadvantage is the limited possibility of decision boundaries, because operations are restricted to thresholding, AND and OR. The more recent iterative Boolean combination proposed by Khreich et al in (Khreich *et al.*, 2010b), extend the approaches in (Tao and Veldhuis, 2009; Barreno *et al.*, 2008) to use not only two, but ten binary combination rules between individual ROC points. ROC based decision-level fusion methods share the advantage that combine points in the ROC space, and matching score normalization is not needed. Although decision and score fusion levels carry the possibility that representational information is lost during combinations, the lower complexity of the combination method and superior performance of the final system usually compensates the drawback.

The following section describes the iterative Boolean combination (IBC) as a state of the art algorithm for the selection and combination of classifiers based on their performance evaluated in the ROC space. This algorithm was employed for classifier combination in most of the methods proposed in this Thesis due to its robustness and accuracy.

1.4.2.1 Iterative Boolean Combination

The iterative Boolean combination (IBC) algorithm is capable of adapting the fusion function and pool of base classifiers according to the most recently acquired data. In this way, the ensemble is adapted to changes in the probability distribution of data in the feature space. Following a learn-and-combine strategy, when new data becomes available a diverse pool of classifiers is generated and combined with previously-learned classifiers. In order to take into account for class imbalance, a validation set with specific imbalance may be employed to represent the expected characteristics of the operational environment. The algorithm for Boolean combination of classifiers proposed by W. Khreich (Khreich *et al.*, 2010b), applies ten Boolean operations shown in Table 1.4 to combine their responses and improve the convex hull.

Table 1.4 Table of truth of the Boolean functions used in Boolean Combination extracted from (Khreich *et al.*, 2010b)

C_a	C_b	$C_a \wedge C_b$	$\neg C_a \wedge C_b$	$C_a \wedge \neg C_b$	$\neg(C_a \wedge C_b)$	$C_a \vee C_b$	$\neg C_a \vee C_b$	$C_a \vee \neg C_b$	$\neg(C_a \vee C_b)$	$C_a \oplus C_b$	$\neg(C_a \oplus C_b)$
0	0	0	0	0	1	0	1	1	1	0	1
0	1	0	1	0	1	1	1	0	0	1	0
1	0	0	0	1	1	1	0	1	0	1	0
1	1	1	0	0	0	1	1	1	0	0	1

Algorithm 1.4 shows the pseudo code of the BC algorithm, which estimates the operations points with highest performance as plotted in the ROC space, according to a validation set.

Then, the Boolean combination of multiple classifiers can be extended following the same approach, with the $SBCM_{ALL}$ strategy. Such an algorithm applies BC_{ALL} to the first pair of ROC curves, and the result is then combined with the third, fourth and so on (see Algorithm 1.5).

Algorithm 1.4: Boolean Combination of classifiers BC_{ALL}

Input : Thresholds of ROC curves, T_a and T_b, D_t^c labels and *skew*
Output : PROCCH and fused responses (Rab) of combined curves
$m \leftarrow |T_a|$
$n \leftarrow |T_b|$
Allocate $F_{2\times m\cdot n}$ `// Temporary results of fusions`
$BooleanFunctions \leftarrow \{a\wedge b, \neg a\wedge b, a\wedge \neg b, \neg(a\wedge b), a\vee b, \neg a\vee b, a\vee \neg b, \neg(a\vee b), a\oplus b, \neg(a\oplus b)\}$
Compute $ROCCH_{old}$ of original curves
for $All\ bf \in BooleanFunctions$ **do**
  for $i = 1,...,m$ **do**
    Convert threshold of the 1st ROC, to responses in R_a
    for $j = 1,...,n$ **do**
      Convert threshold of the 2nd ROC to responses in R_b
      Combine responses with bf, and produce R_c
      Compute (fpr,tpr) using R_c, labels
      Push (fpr,tpr) onto $\mathscr{F}$ (temporary results of fusions)
  Compute $ROCCH_{new}$ (Convex Hull) of $F \cup ROCCH_{old}$
  $s^*_{global} \leftarrow (T_{a_x}, T_{b_y}, bf)$ `// to be used during operations`
  Store responses of these emerging points into R `// to be used with` BCM_{ALL} `and` IBC_{ALL}
  $ROCCH_{NEW} \leftarrow ROCCH_{OLD}$ `// Update ROCCH`
Return: $ROCCH_{NEW}, R, s^*_{global}$

Algorithm 1.5: Boolean Combination of multiple classifiers BCM_{ALL}

Input : Thresholds of ROC curves, $[T_1,...T_K]$ and D_t^c labels
Output : ROCCH and responses of the combination (R)
$[ROCCH_1, R_1] \leftarrow BC_{ALL}(T_1, T_2, labels, skew)$
for $k = 3,...,K$ **do**
  $[ROCCH_{k-1}, R_{k-1}] \leftarrow BC_{ALL}(T_{k-2}, T_k, labels, skew)$
Return : $ROCCH_{K-1}, R_{K-1}$ and stored tree of selected responses/thresholds fusions along with their corresponding fusion functions

Also the iterative combination of such curves can be extended, using the ROC-AUC as the performance measure to maximize given a determined skew (Algorithm 1.6).

Algorithm 1.6: Iterative Boolean Combination IBC_{ALL}

Input : Thresholds of ROC curves, $[T_1,...T_K]$ and D_t^c labels
Output : PROCCH and fused responses (R)
$[ROCCH_{OLD}, R_{OLD}] \leftarrow BCM_{ALL}([T_1...T_K], labels)$
while $(AUC(ROCCH_{NEW}) \geq AUC(ROCCH_{OLD}) + \varepsilon)$ *or* $(iterations \leq maxIter)$ **do**
  $[ROCCH_{NEW}, R_{NEW}] \leftarrow BCM_{ALL}(R_{OLD}, [T_1...T_K], labels)$
Return: $ROCCH_{NEW}, R_{NEW}$ and stored tree of selected responses/thresholds along with their corresponding fusion functions

1.4.3 Ensembles for Class Imbalance

The algorithms designed for environments with changes in the probability distribution of data in general, and particularly in the class priors, can be categorized according to the use of a mechanism to detect changes in prior probabilities (Ditzler and Polikar, 2013). Approaches with active detection of changes in prior probabilities seek explicitly to determine whether and when a change has occurred in the prior probability before taking a corrective action (Radtke *et al.*, 2013a,b; Ditzler and Polikar, 2013). Conversely, approaches with passive change detection assume that a change may occur at any time, or is continuously occurring, and hence the classifiers are updated every time new data becomes available (Oh *et al.*, 2011; Ditzler and Polikar, 2013). The rest of this section describes representative approaches of passive and active ensembles for changing priors.

1.4.3.1 Passive Approaches

Passive *ensemble-based* methods for class imbalance can be categorized in cost-sensitive ensembles, boosting-based, bagging-based and hybrids (Galar *et al.*, 2011). In cost-sensitive approaches, the combination of classifiers (i.e. weights) is designed to consider the cost of class independent errors. Examples of these approaches include the AdaCost, CSB, RareBoost, AdaC1, AdaC2 and AdaC3 algorithms (Fan *et al.*, 1999; Wu, 2012). Boosting-based ensembles include techniques that use data preprocessing embedded into boosting algorithms. These methods bias the data distribution towards the minority class before the classifier generation step. Examples of these approaches are the Learn++.CDS, Learn++.NIE, SMOTEBoost, MSMOTEBoost, RUSBoost and DataBoost-IM algorithms (Ditzler and Polikar, 2013, 2010). Bagging-based ensembles integrate bagging with data preprocessing techniques, and hence, they do not require to update any kind of weights. These techniques address the class imbalance by the way they collect the training samples, using oversampling and/or undersampling techniques to generate training sets of different sizes. Examples of these techniques are the OverBagging, UnderBagging, UnderOverBagging and Imbalanced IVotes (Wang and Yao, 2009; Barandela *et al.*, 2003). Finally, hybrid ensembles combine a pre-processing tech-

nique with a bagging and a boosting technique. Techniques in this category are also called exploratory undersampling, and basically include EasyEnsemble and BalanceCascade (Liu *et al.*, 2009).

Although the aforementioned methods account for class imbalance through adaptation every time new reference samples become available, they are passive since they do not perform an estimation of the imbalance before adaptation. The advantage of passive approaches lies in the avoidance of false positive and false negative change detections, at the cost of the increased complexity of continuous adaptation.

1.4.3.2 Active Approaches

Active methods for adaptation to class imbalance employ a mechanism to estimate the class priors of the input data, and adapt the algorithm to the estimated class proportions when a change occurs. Hence, these approaches avoid the assumption of continuous changes and the complexity of continuous adaptations, with the potential disadvantage of false positive and false negative change detections. Several examples of active approaches that employ ensembles for classification in imbalanced environments appear in literature (Radtke *et al.*, 2013a,b; Wang *et al.*, 2013a). In general, passive approaches for changing imbalance can be modified by adding a mechanism to detect changes in prior probabilities. Some examples of such mechanisms are based in Hellinger distance (Radtke *et al.*, 2013b), Kullback Leibler divergence (du Plessis and Sugiyama, 2012), or accounting for class-specific performance measures like *recall* (Wang *et al.*, 2013a,b).

A recently proposed active approach employed in face recognition in video surveillance is the skew-sensitive Boolean combination (SSBC), which estimates the imbalance using the Hellinger distance between the distributions of validation data and the most resent unlabeled operational samples (Radtke *et al.*, 2013b).

1.4.3.3 Skew-Sensitive Boolean Combination

The IBC strategy efficiently integrates the responses of multiple diversified classifiers in the ROC space, yet the impact on performance of imbalanced data distributions is difficult to observe from ROC curves. The Skew-Sensitive Boolean Combination (SSBC) technique exploits the Precision-Recall Operating Characteristic (PROC) space, leading to a higher level of performance (Radtke *et al.*, 2013b). A set of BCs of base classifiers is initially produced with imbalanced reference data in the PROC space, where each BC curve corresponds to different level of imbalance (a growing number of non-target samples versus a fixed number of target ones). Then, during operations, the closest adjacent levels of class imbalance are periodically estimated using the Hellinger distance between the data distribution of inputs and that of imbalance levels, and used to approximate the most accurate BC of classifiers from operational points of these curves. In this manner, the ensemble is capable of on-line adaptation of the fusion function to the most recent operational imbalance.

During training, SSBC assumes that a diversified pool of binary classifiers $\mathscr{P} = \{p_1, ..., p_n\}$, and operates at the combination level to take advantage of the diversity of opinions in the ensemble. To do that, validation data with different levels of imbalance is used to estimate the operations points of the Boolean combination function (covering the whole ROC space). Two validation sets with that imbalances, the first (`OPT`) employed to estimate the operational imbalance, and the other (`VAL`) to select the operation point with the propper estimated imbalance. During operations, the imbalance is estimated using the Hellinger distance, and the operation points are selected from the predefined imbalances. The known levels of class imbalance used by the approach form the set $\Lambda = \{\lambda^{bal} = 1:1, ..., \lambda^{max}\}$. A subset of class imbalances $\Lambda_{BC} \subset \Lambda$ is selected from Λ to optimize a subset of BCs E. The subset of imbalances Λ_{BC} should contain evenly distributed intermediate class imbalance levels between the minimum λ^{bal} and the maximum level of imbalance λ^{max} inclusively. The sets `OPT` and `VAL` are generated from imbalanced reference data that follows λ^{max}. Different data sets with the levels of class imbalance defined in Λ, in which the amount of target samples remains fixed, while the amount of non-target samples are added to the set through random under sampling.

The classification system operates by receiving streams of operational feature vectors corresponding to facial regions detected in video. The operational histogram opd corresponding to these operational samples is accumulated over time, and the closest level of class imbalance $\lambda^* \in \Lambda$ is estimated by comparing opd to the data sets in OPT using the Hellinger distance. Recall that each data set in OPT follows a class imbalance from Λ. The estimated operational class imbalance λ^* corresponds to the imbalance of the closest set in OPT to opd in terms of Hellinger distance.

The Hellinger distance is a measure of similarity between two sets and is defined as follows. Given an unlabeled dataset $U = \{(\mathbf{a}^n), n = 1,...,N\}$ and a labeled validation dataset $V = \{(\mathbf{a}^m, l^m), m = 1,...,M\}$, the Hellinger distance between these two sets can be computed according to

$$HD(V,U) = \frac{1}{n_f} \sum_{f=1}^{n_f} HD_f(V,U), \tag{1.9}$$

where the feature-specific Hellinger distance is given by

$$HD_f(V,U) = \sqrt{\sum_{i=1}^{b} \left(\sqrt{\frac{|V_{f,i}|}{|V|}} - \sqrt{\frac{|U_{f,i}|}{|U|}} \right)^2}, \tag{1.10}$$

where n_f is the number of features, b is the number of bins used to construct the feature-specific histogram representation of the probability density functions of the datasets. $|U|$ is the number of samples in U and $|U_{f,i}|$ is the number of samples whose feature f belongs to the bin i, similarly with $|V|$ and $|V_{f,i}|$ for the validation set V.

The operational imbalance λ^* estimated using the Hellinger distance is used to select the BC that corresponds to that imbalance, and in the case λ^* is not available on Λ_{BC}, the BCs for the two closest imbalances are merged, and the convex hull is estimated (see Algorithm 1.7). In Algorithm 1.7, the imbalanced sets of reference data OPT and VAL allow to select subsets of target and non-target samples with different imbalances up to a maximum pre-determined λ^{max}. The subsets $\texttt{opt}^* \in \texttt{OPT}$ and $\texttt{VAL}^* \in \texttt{VAL}$ are generated according to the desired imbalance λ^*.

The whole set of target samples is maintained, and non-target samples are randomly selected (random undersampling) to obtain the desired imbalance.

Algorithm 1.7: SSBC technique for adapting BC for a new class imbalance level λ^*

Input : set of BCs E, set of class imbalance levels Λ_{BC}, data sets `OPT` and `VAL`, the estimated class imbalance $\lambda^* \in \Lambda$ and the target fpr.
Output : Operations point op for the target fpr.
if $\lambda^* \in \lambda_{BC}$ **then**
  $E^* = E_{\lambda^*}$
else
  Select $\lambda^i, \lambda^j \in \Lambda_{BC}$, such that $\lambda^i < \lambda^* < \lambda^j$
  Select $\texttt{opt}^* \in \texttt{OPT}$, following λ^*
  $E^* = ROCCH(E_{\lambda^i} \cup E_{\lambda^j}, \texttt{opt}^*)$
Select $\texttt{val}^* \in \texttt{VAL}$, following λ^*
Select $op \in E^*$ for the target fpr with val^*

The strength of the SSBC algorithm lies in the adaptive selection of suitable fusion functions (ROC operations points) according to the estimated operational imbalance. However, this technique assumes that the generation of a pool of classifiers, where each classifier is trained using balanced target and non-target data, and provide enough diversity of opinions to discriminate when input operational data is imbalanced. Another issue is related to the precision of the method used by SSBC to estimate the class imbalance is limited by the amount and sampling strategy used to create the set of imbalances Λ.

1.4.4 Challenges on Adaptive Ensembles for Class Imbalance

Exploiting imbalance to adapt a classifier system has been studied in literature, and is a consequent option regarding the imminent imbalance in face based video surveillance. Although the algorithms like SSBC have successfully used imbalanced validation data to update an ensemble fusion function to the operational imbalance, two issues are still to be addressed in practice. The first is related to the source of diversity of opinions among experts, where classifiers may be trained on data with different imbalances and complexities. In this way, the base classifiers trained on diverse levels of imbalance would provide increased useful diversity in

the ensemble. Even more, training imbalance specific classifiers on data with different complexities would provide even more diversity, leading to a more accurate and robust ensemble under such an imbalanced environment.

The second issue is related to the resolution needed to reliably estimate the operational imbalance. For example, SSBC estimation relies on the measurement of the Hellinger distance between the histogram representation of a set with the most recent operational samples and validation sets with pre-defined imbalance levels (Λ). If the operational imbalance is not considered in the set Λ, the combination functions corresponding closest adjacent imbalances are considered, but the exact level of imbalance is never estimated. More accurate candidate quantification methods like HDx and HDy may be used, where all the validation samples are employed for a more precise estimation, avoiding the subsampling requirement.

1.5 Measuring Classification Performance

In this Thesis a modular architecture is assumed, composed of an ensemble of 2-class classifiers for each individual, and the performance for these detectors is measured as in verification problems – e.g. using binary decision spaces. The measurement of performance for these classifiers is explored in this section.

Probably the most commonly used measure of performance when evaluating classifiers is the accuracy, or its complement, error rate. It is well known that a classifier that produces less mistakes is preferable. However it is sometimes convenient to use other measures that focus on a determined type of errors (e.g. *false positive rate* or *false negative rate*). The confusion matrix is widely used when evaluating binary classifiers (Table 1.5).

Table 1.5 Confusion matrix for a binary classifier

	Actual positive	**Actual negative**
Predicted positive	True Positives (TP) Positives correctly classified	False Positives (FP) Negatives incorrectly classified
Predicted negative	False Negatives (FN) Positives incorrectly classified	True Negatives (TN) Negative correctly classified

From the confusion matrix several measures are derived (See Table 1.6) with different characteristics and each of them is useful under specific conditions.

Table 1.6 Common measures derived from the confusion matrix

$sensitivity = recall =$ True Positive Rate $= tpr = \frac{TP}{TP+FN}$	False Positive Rate $= fpr = \frac{FP}{FP+TN}$
True Negative Rate $= tnr = \frac{TN}{FP+TN} = 1 - fpr$	False Negative Rate $= fnr = \frac{FN}{TP+FN} = 1 - tpr$
$Accuracy = \frac{TP+TN}{P+N}$	$precision = \frac{TP}{TP+FP}$
$F - measure = \frac{2}{1/precision+1/recall}$	$specificity = 1 - fpr$

Using distinct pairs of these measures lead to different 2-dimensional performance spaces, which present distinct properties. As shown in Figure 1.7, one point (classifier) represented in the ROC space dominates another if it is above and to the left: has a higher tpr and a lower fpr.

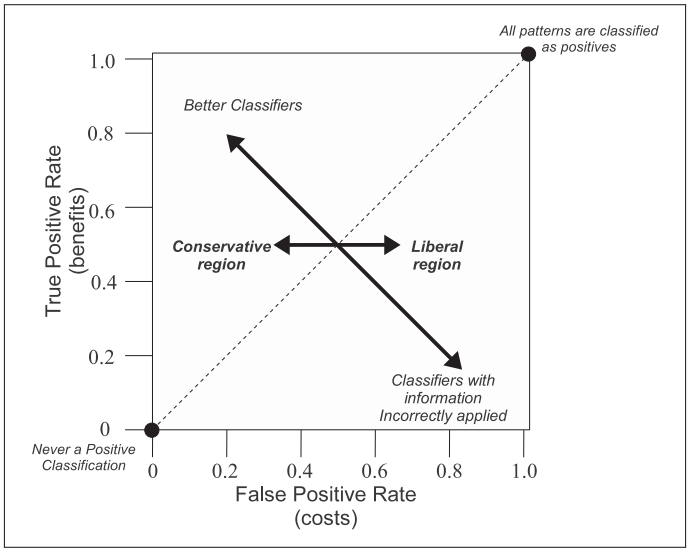


Figure 1.7 ROC space and its different regions
extracted from (Flach, 2004)

ROC curves are insensitive to changes in class distribution (proportion of positive to negative instances). An operating point on the curve is a specific combination of misclassification costs and class distributions (Fawcett, 2006). Some limitations of visual inspection using ROC curves are stated in (Drummond and Holte, 2006). In ROC space, it is not possible to know

what is the performance (expected cost) of a classifier, neither the difference in performance between two classifiers. It is also not possible to know for what misclassification costs and class probabilities is the difference in performance between two statistically significant classifiers. Besides, in ROC space it is not possible to know for what misclassification costs and class probabilities a given classifier outperform the trivial classifier that assigns all samples to the same class.

The space of Cost Curves maps each point in ROC space to a line as shown in Figure 1.8. The slope of the segment of the convex hull in ROC space, that connects two points (fpr_1, tpr_1), (fpr_2, tpr_2), is given by

$$slope = \frac{tpr_1 - tpr2}{fpr_1 - fpr_2} = \frac{p(-)C(+|-)}{p(+)C(-|+)} \tag{1.11}$$

where $p(a)$ is the probability of a given sample to be positive ($a = +$) or negative ($a = -$), and $C(a|b)$ is the cost when a sample a is classified as b (Drummond and Holte, 2006).

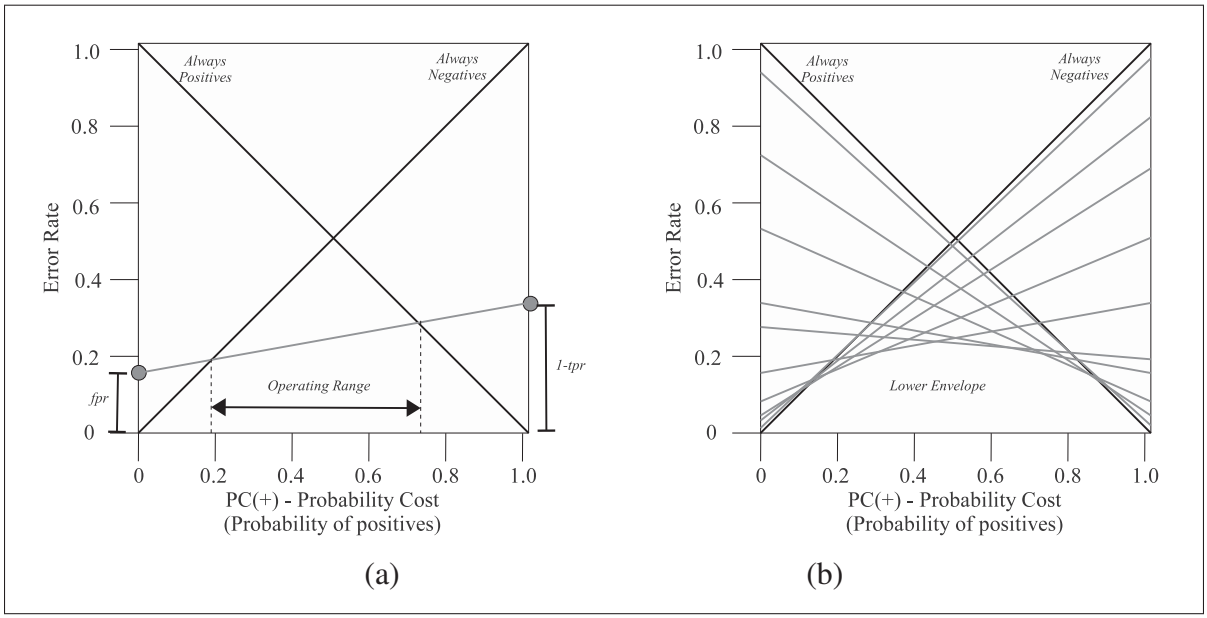


Figure 1.8 Cost curves space, one point in ROC space maps to a line in cost curves space (a), and varying a threshold generates a set of lines (b) extracted from (Flach, 2004)

Cost curves visually support several crucial types of performance assessments that cannot be done easily with ROC curves, such as showing confidence intervals and visualizing statistical significance on difference in performance of two classifiers. This curves are specifically designed for a specific performance measure: the expected cost. The area under the cost curve is the expected cost of the classifier assuming all possible probability-cost values are equally likely. A perfect classifier is represented as a horizontal line from the point $(0,0)$ to $(0,1)$, representing zero cost for any probabilities. For any X point, the corresponding Y points represent the expected costs of the classifiers. In ROC space the convex hull contains the set of lowest-cost classifiers. In cost space, the lower envelope represents this set (Fawcett, 2004).

A disadvantage of ROC analysis is that given its invariance to variations in class priors, it hides an important factor of evaluation in imprecise environments, where misclassification costs can not be specified exactly, and class priors may not be reflected by the sampling. This is also equivalent to the lines projected to the cost space, where to select operating points (lines), misclassification may be specified. Even more, priors in imprecise environments may vary continuously, and optimal decision threshold selection may be ill-defined. In this cases, *precision-recall* space remains sensitive to performance on each class (Landgrebe *et al.*, 2006; Davis and Goadrich, 2006).

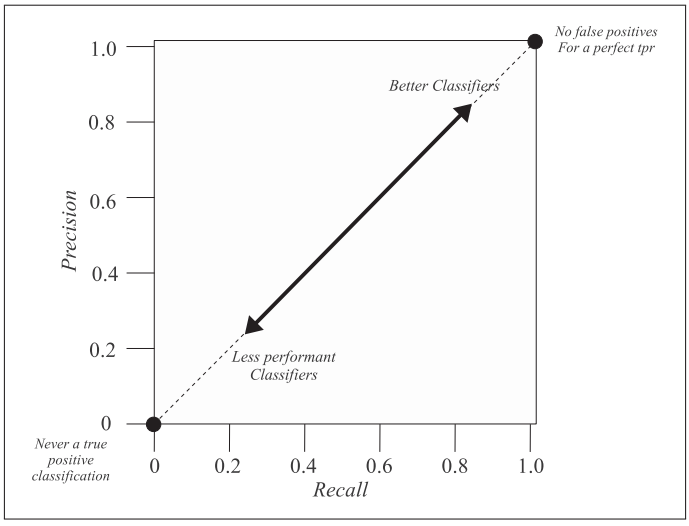


Figure 1.9 PROC ($precision - recall$) space

P-ROC graphs have been used in applications where the number of negative samples is many orders of magnitude greater than positives, and probably this ratio increases steadily as the number of samples increase (Fawcett, 2004). P-ROC curves are not insensitive to changes in class distribution, which is desirable in applications where high skew is expected (e.g. 10^6 negatives for 1 positive sample).

1.6 Summary of Overall Challenges

Video-based FR systems employed in video surveillance face numerous challenges related to the time and spatial variations in capture conditions. Changes due to the natural ageing of people and variations in capture conditions and camera interoperability induce gradual and abrupt changes in the facial appearance of enrolled individuals during operation, and hence, in the classification environment. Facial models are designed a priori with a limited amount of reference faces that are often captured under controlled conditions at enrollment time, and therefore loose their representativeness over time. The performance of a system for video-to-video FR is significantly degraded due to these factors.

Whereas modular classification strategies may allow to reduce the complexity of the problem faced by multi-class classifiers to find multiple decision frontiers, they add the robustness of ensemble techniques. Adaptive multiple classifier systems (MCS) capable of incremental learning allow to update the facial models with new reference facial captures, providing the posibility to maintain performance after the classification environment changes. However, the requirement of manual acquisition and labeling of the new reference data is still an issue (it is costly or unfeasible in practice). In this Thesis, the strategy to address this problem consists in an adaptive system inspired in semi-supervised learning, but employing video-to-video strategies for self-update. However, self-updating stragegies affect a trade-off between self-adaptation and accuracy of facial models.

Finally, the proportions of target and non-target individuals in face re-identification are imbalanced, and these proportions also change over time, affecting the performance of the classifica-

tion systems employed for matching. Although the SSBC algorithm was proposed to estimate target vs. non-target proportions periodically during operations, there are some challenges that arise from using SSBC in practical VS applications. For instance, SSBC is typically used to combine a pool of classifiers designed on balanced data, ignoring the diversity that can be provided by employing several imbalance levels to train base classifiers. And the estimation of operational imbalance depends on the availability of the validation set with a similar imbalance, making it difficult to select the set of validation imbalances.

CHAPTER 2

PARTIALLY-SUPERVISED LEARNING FROM FACIAL TRAJECTORIES FOR FACE RECOGNITION IN VIDEO SURVEILLANCE

Miguel De-la-Torre[1,2], Eric Granger[1], Paulo V. W. Radtke[1], Robert Sabourin[1], Dmitry O. Gorodnichy[3]

[1] Laboratoire d'imagerie de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montréal, Canada

[2] Centro Universitario de Los Valles, Universidad de Guadalajara, Ameca, México

[3] Science and Engineering Directorate, Canada Border Services Agency, Ottawa, Canada

Paper published in the journal "Information Fusion", from Elsevier, June 2014

ABSTRACT

Face recognition (FR) is employed in several video surveillance applications to determine if facial regions captured over a network of cameras correspond to a target individuals. To enroll target individuals, it is often costly or unfeasible to capture enough high quality reference facial samples a prior to design representative facial models. Furthermore, changes in capture conditions and physiology contribute to a growing divergence between these models and faces captured during operations. Adaptive biometrics seek to maintain a high level of performance by updating facial models over time using operational data. Adaptive multiple classifier systems (MCSs) have been successfully applied to video-to-video FR, where the face of each target individual is modeled using an ensemble of 2-class classifiers (trained using target vs. non-target samples). In this chapter, a new adaptive MCS is proposed for partially-supervised learning of facial models over time based on facial trajectories. During operations, information from a face tracker and individual-specific ensembles is integrated for robust spatio-temporal recognition and for self-update of facial models. The tracker defines a facial trajectory for each individual that appears in a video, which leads to the recognition of a target individual if the positive predictions accumulated along a trajectory surpass a detection threshold for an ensemble. When the number of positive ensemble predictions surpasses a higher update threshold, then all tar-

get face samples from the trajectory are combined with non-target samples (selected from the cohort and universal models) to update the corresponding facial model. A learn-and-combine strategy is employed to avoid knowledge corruption during self-update of ensembles. In addition, a memory management strategy based on Kullback-Leibler divergence is proposed to rank and select the most relevant target and non-target reference samples to be stored in memory as the ensembles evolves. For proof-of-concept, a particular realisation of the proposed system was validated with videos from Face in Action dataset. Initially, trajectories captured from enrollment videos are used for supervised learning of ensembles, and then videos from various operational sessions are presented to the system for FR and self-update with high-confidence trajectories. At a transaction level, the proposed approach outperforms baseline systems that do not adapt to new trajectories, and provides comparable performance to ideal systems that adapt to all relevant target trajectories, through supervised learning. Subject-level analysis reveals the existence of individuals for which self-updating ensembles with unlabeled facial trajectories provides a considerable benefit. Trajectory-level analysis indicates that the proposed system allows for robust spatio-temporal video-to-video FR, and may therefore enhance security and situation analysis in video surveillance.

2.1 Introduction

In video surveillance applications, automated face recognition (FR) systems are increasingly employed to match facial regions of interest (ROIs) captured across a network of video cameras to individuals of interest enrolled to the system. These applications range from watch-list screening, which involves still-to-video FR, to person re-identification (for search and retrieval), which involves video-to-video FR. Regardless, systems for FR in video surveillance (FRiVS) must operate under semi- and unconstrained capture conditions, where scale, pose, occlusion, blur/resolution, expression and illumination vary over time.

A facial model used for matching may be defined as a set of one or more reference samples (for a template matching system), or a statistical model estimated through training with reference samples (for a neural or statistical classification system). In video-to-video FR, reference sam-

ples extracted from ROIs captured in video streams are employed to design of facial models, integrating time and space information in facial models (Li and Wechsler, 2005; De-la Torre *et al.*, 2012a). In still-to-video FR, reference samples are extracted from one or more still images.

In video surveillance, individuals in a scene may be tracked, and the facial ROIs captured in videos that correspond to different individuals may be regrouped over multiple frames for robust spatio-temporal recognitions (Matta and Dugelay, 2009). Tracking information can, for instance, be used to record a complete trajectory[1], from the arrival of individual in the scene until he leaves. Predefined thresholds have been applied to matching scores and image quality measurements to produce overall decisions based on the consecutive ROIs (Despiegel *et al.*, 2012). In addition, the sum rule has been applied over the matching scores produced by ROIs in a trajectory (Ekenel *et al.*, 2010). Tracking information as also been used to model the joint posterior distribution of the motion and identity for the individual in the scene (Zhou *et al.*, 2003).

This chapter concerns system for video-to-video FR, where facial models for matching are defined as a statistical model. Facial models are usually designed during enrollment, ideally using several high quality reference ROIs captured for the target individual under controlled conditions. In video-to-video FR, these reference ROIs are extracted along one or more reference trajectories. This requirement is rarely fulfilled in practical applications, and enrollment of individuals often relies on a limited number of lower quality ROIs. FR performance tends to decline since facial models are not representative of the faces to be recognized during operations. Both abrupt and gradual changes in capture conditions (due to, e.g., aging and variations in pose and lighting) also lead to a decline in FR performance due to a growing divergence between these facial models and faces captured during operations. Several adaptive classifiers have been proposed in literature for supervised incremental learning of labeled samples (De-la Torre *et al.*, 2012a; Polikar *et al.*, 2001; Singh *et al.*, 2010; Connolly *et al.*, 2012). These can be

[1] A *facial trajectory* is defined as a set of ROIs (isolated through face detection) that correspond to a same high quality track of an individual across consecutive frames.

used to update facial models after enrollment, as new reference data becomes available, allowing to maintain or increase matching performance. Adaptive multiple classifier systems (MCS) have been successfully applied for FRiVS (De-la Torre *et al.*, 2012a; Pagano *et al.*, 2012). In these systems, the facial model of each individual is encoded using an ensemble of 2-class classifiers or detectors (EoD), trained to discriminate between samples of a target individual and non-target individuals.

An issue with the supervised update of classifiers is the analysis and extraction of labeled reference samples from operational videos. A domain expert must isolate target faces manually or semi-automatically in video surveillance footage, which involves undesirable costs and delays. Instead of relying on a human expert, the system may self-update face models with operational videos. Several semi-supervised learning approaches have been proposed to update biometric models using a combination of labeled and unlabeled samples (Rattani, 2010; Roli and Marcialis, 2006; Franco *et al.*, 2010). In the area of adaptive biometrics, two representative approaches for semi-supervised learning are the self-update and co-update techniques (Roli *et al.*, 2007). The first applies an update threshold (higher than the detection threshold) to each matching scores to select input biometric samples as new templates, and the second seeks corroboration of scores from two or more matchers for cross-updating.

To the authors' knowledge, a FR system that allows for self-updating facial models in video surveillance applications has not been proposed in literature. An issue encountered with self-updating is the reliable selection of operational samples from the target individual to adapt facial models. A high level of confidence is required to avoid updating models with non-target data. In contrast, a facial model should also be adapted with a diversified set of reference samples to improve the generalization performance. Given an adaptive MCS proposed in (De-la Torre *et al.*, 2012a; Pagano *et al.*, 2012), information from a face tracker and individual-specific ensembles may be integrated to provide a variety of high confidence reference samples.

In video surveillance, an abundance of reference samples may be extracted from non-target facial trajectories acquired in the scene during routine system operation. Two databases may

be formed with samples extracted (1) from trajectories of other individuals of interest besides the target individual (known as the cohort model, CM), and (2) from unknown people appearing in scene (known as the universal model, UM) (Li and Wechsler, 2005; De-la Torre *et al.*, 2012a; Pagano *et al.*, 2012; Merati *et al.*, 2010). This imposes the need to sub-sample non-target data in order to design accurate facial models, using an ensemble of 2-class classifiers. Moreover, adaptive MCSs require reference data to be stored in memory for validation (De-la Torre *et al.*, 2012a; Connolly *et al.*, 2012). Practical memory limitations impose the need for a method to rank and select the most relevant validation samples for each individual (EoD).

In this chapter, an adaptive MCS is proposed for video-to-video FR in semi- and unconstrained video surveillance environments. Within the adaptive MCS, an EoD encodes and updates the facial model of each individual of interest. This novel system allows for spatio-temporal recognition and self-update of facial models based on high-confidence trajectories. During operations, a face tracker defines facial trajectories for different individuals that appear in a video. Track ID numbers are integrated with predictions of individual-specific ensembles at a decision-level for enhanced video-to-video FR. The proposed system relies on tracker quality to regroup ROIs into facial trajectories, and applies a double thresholding scheme to curves produced by accumulating positive EoD predictions for a trajectory. An individual of interest is recognized if the number of positive predictions accumulated over some time window of a trajectory surpass a *detection* threshold for an EoD.

A second (higher) *update* threshold is applied to select high-confidence trajectories that are suitable for self-updating a facial model. If the number of positive predictions surpasses this threshold for an EoD, then all samples extracted from the target ROIs of the trajectory are combined with non-target samples (selected from the CM and UM) to update the corresponding face model. Since a trajectory may contain target ROIs that were incorrectly classified by the EoD, facial models are adapted with a diversified set of reference samples that may refine the decision boundary between target and non-target distributions, and thereby improve the generalization performance. A sub-sampling technique based on condensed nearest neighbor (CNN) (Hart, 1968) is employed to select non-target samples along this boundary. The data for

EoD update is comprised of diverse facial regions associated with target and non-target trajectories, and is employed to generate a new pool of 2-class classifiers, and to update the fusion function of the user specific EoD. To avoid issues related to knowledge corruption in incremental learning classification systems, the self-update of EoD employs a learn-and-combine strategy (De-la Torre *et al.*, 2012a). Finally, a long term memory (LTM) is maintained over time with a fixed number of reference validation samples per individual. A memory management strategy based on the Kullback-Leibler (KL) divergence criteria (Kachites McCallum and Nigam, 1998) is proposed to rank and select the most relevant target and non-target reference samples samples. This criteria seeks to preserve the highest relative entropy of ensemble over time. In other words, the KL divergence becomes higher for samples that contain a higher level of information according to the knowledge previously acquired by the individual specific EoD.

Video sequences from the Carnegie Mellon University Face in Action (FIA) dataset for video FR was used for proof-of-concept validation. Video sequences were captured from 180 subjects with an array of 6 cameras over three sessions separated by a three-month interval. In this dataset, video of individuals were captured under semi-controlled conditions in a security check point scenario. When a sequence is presented to the proposed system during operations, trajectories are employed for spatio-temporal recognition, and high-confidence trajectories are used for self-update. Three levels of performance evaluation are considered – transaction-based analysis (in the ROC and *precision-recall* spaces), subject-level analysis (Doddington zoo characterization), and trajectory-based analysis (of the overall system for video sequences).

This chapter is organized as follows. Sections 2.2 and 2.3 provide a brief overview of techniques employed for FRiVS and adaptive biometrics, respectively. The adaptive MCS proposed for self-update from facial trajectories is described in Section 2.4, including specialized individual-specific strategies for management of reference data, for fusion of tracking and classification responses, and for self-update of facial models (EoDs). Section 2.5 describes the experimental methodology – protocol, video data set and measures used in performance evaluation. Finally, results are presented and discussed in Section 2.6.

2.2 Video-to-video Face Recognition

Assume that video streams are captured using one or more video cameras (see Fig. 2.1). The segmentation process isolates the facial regions of interest (ROIs) from successive frames, and discriminative features are extracted to represent faces for tracking (vector **b**) and classification (vector **a**). A new track is typically initialized when an emergent face is captured far from others, and is defined over consecutive frames using the state of the facial region being tracked **x** (appearance, scale, position, track number, etc.) and a vector of tracker-specific features **b**. Classification features extracted from each ROI (vector **a**) are often image-based (using e.g., Local Binary Patterns) or pattern recognition-based (using e.g., Principal Component Analysis). The tracking module follows the movement or expression of distinct faces across video frames, while the classification module matches ROIs captured in video to the system's facial models. Finally, the decision fusion combines track **ID**s and classification scores **s** in order to predict it target individuals appear before a camera.

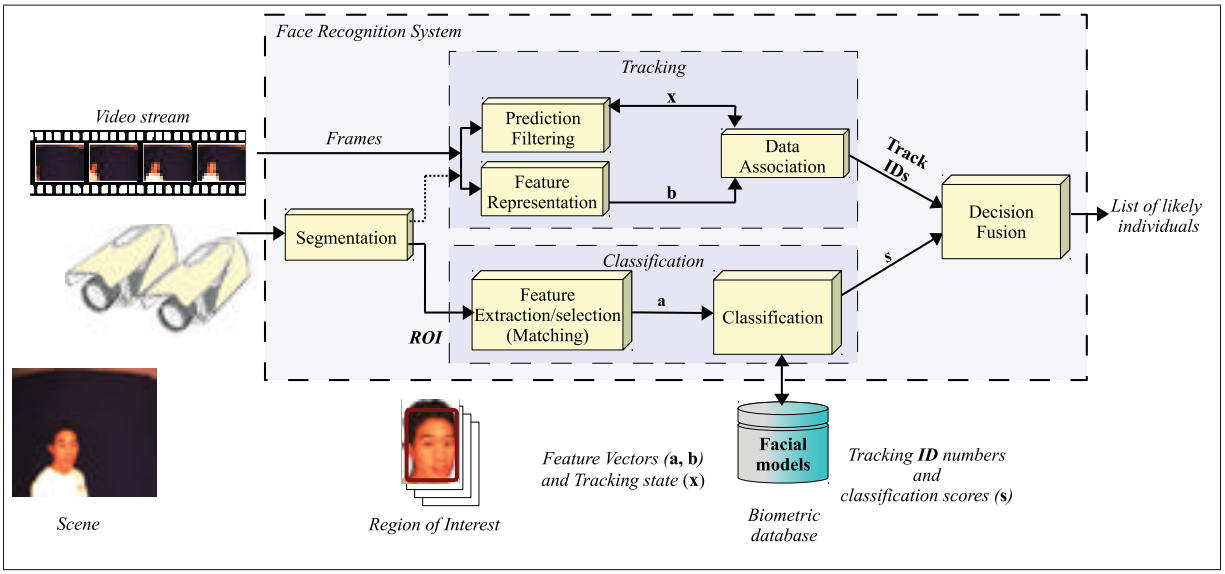


Figure 2.1 Block diagram of a system for video face recognition

2.2.1 Face Tracking

Facial tracking (FT) techniques allow to follow the movement of each of individual and to regroup facial regions of a same person (without knowing his identity). The input of the tracker is the stream of frames acquired with video cameras, and the initial face ROIs to be tracked, while the output defines as a set of facial regions with the same **ID** for which the track has high tracking quality Q_T. Note that only the first ROI in a trajectory (ROIs used for classification) may be equivalent in a track (state of facial regions from the tracker) (Yilmaz *et al.*, 2006).

The basic tracking steps are face representation, prediction filtering and data association. In face representation, the tracked facial region is represented with distinctive features (tracking feature vector **b**) in order to allow tracking from one frame to the next. Commonly used features are color histogram, skin color probability map and active contours, just to mention a few. Predicting the next state with Kalman and Particle filters seeks the new state **x** (appearance, scale, location, and/or velocity, etc.) of the facial region to be tracked in the current frame, based on the information in the previous frames and some underlying model for state transitions. The objective of the prediction filtering is to avoid drift and reduce the search space by using a probability framework, although some methods perform data association heuristically instead (e.g. Mean-shift and Cam-shift). Finally, in the data association step, the tracker associates a feature vector of the facial region extracted from the previous frame with the feature vector in the current frame. Tracking methods are categorized according to the type of descriptor used for face representation: holistic, contour-based, and hybrid information. Most face-tracking methods in literature rely on holistic representations due to their robustness.

2.2.2 Specialized Classification Architectures

In the literature, FR in video surveillance (FRiVS) is addressed as an open set problem, considering that the number of individuals of interest is highly outnumbered by other persons in the scene. Multi-class classifiers have been used, which apply a rejection threshold for unknown individuals. A multi-class classifier designed for video FR is the Open Set TCM-kNN (Li

and Wechsler, 2005). It uses transductive inference to produce a classification score based on randomness deficiency. Tax and Duin also proposed a technique to combine one-class classifiers in a multi-class classifier. Their heuristic allows to adjust a class-specific outlier rejection threshold, and combine non-generative class models (Tax and Duin, 2008).

Similarly, modular architectures with one detector per individual have been proposed to address the problem with individual-specific 1- or 2-class classifiers. The convenience of these modular approaches has been widely studied in the literature, setting individual- (or user-) independent parameters (Jain and Ross, 2002). For instance, the approach proposed by Kamgar and Parsi, that identifies the decision region(s) in the feature space for each individual face by training a dedicated feed-forward neural network for each individual of interest (Kamgar-Parsi *et al.*, 2011). Another example is the SVM-based modular system proposed by Ekenel et al., applied to a visitor interface scenario (Ekenel *et al.*, 2010).

Finally, modular approaches have been extended to train an ensemble of classifiers per individual. An example of such a system is the ensemble of detectors (EoD) designed for each person in a watch list. Non-target samples are retrieved from the CM (database maintained with trajectories from non-target individuals of interest) and the UM (database with training samples from unknown people appearing in scene). Base classifiers are co-jointly trained using a training strategy based on DPSO. It allows for the generation of a diversified pool of ARTMAP neural networks, and trained detectors are then selected and combined using Boolean combination (BC) (Pagano *et al.*, 2012).

2.2.3 Decision Fusion

Approaches for FR in video can be categorized according to those that neglect temporal information and those that propose strategies to exploit it. Algorithms that neglect temporal information have been proposed for still image recognition, and exploit only physiological information on the face. Examples of these approaches include Eigenfaces, Fisherfaces and Active Appearance Models. Alternatively, approaches that exploit temporal information present the

advantage of increased contextual knowledge and data in video, allowing the use of physiological and behavioral information. Discriminant analysis of facial optical flow, Hidden Markov Models (HMMs), and the sequential importance sampling (SIS) algorithm are just some approaches in this category (Matta and Dugelay, 2009).

Spatio-temporal approaches for FR merge spatial information (e.g. face appearance) with the sequential variations presented over time (e.g behavior). Zhang and Martinez use probabilities accumulated by matching ROIs to the individual-specific Gaussian mean estimated from gallery reference samples, and normalize to produce posterior probabilities. This temporal analysis is independent of the matching or tracking algorithm (Zhang and Martinez, 2004). Liu and Chen used HMMs to model the appearance and dynamics of a person, obtaining high confident results on sequences that were then used to adapt the models. A potential problem with the modeling of probability distributions of the motion is the assumption that the movement will be very similar, regardless of the new scenario (Liu and Cheng, 2003). Accumulating classification responses over time eliminates the assumption, and still takes into account the time information. For instance, the work of Ekenel et al. evaluates a video-to-video FR system for individuals entering into a room, which progressively combines confidence scores of the matchers using a sum rule over the full sequences to estimate the identity in video (Ekenel *et al.*, 2010). In their approach, they use a k-NN classifier on a DCT representation of face images, and use min-max normalization on the distance-based output scores, and then compare their proposed approaches: distance-to-model, distance-to-second-closest and a combination of both. *Score* and *quality* driven fusion methods were used to combine responses from frames in video sequences, within a border control system (Despiegel *et al.*, 2012). In the first method, matching scores are compared to a predetermined threshold, whereas the second compares the intrinsic quality of the image intrinsic to the predefined threshold. Finally, a joint sparse representation has been used to simultaneously take into account correlations and coupling information among video frames (Chen *et al.*, 2013). Sub-dictionaries for distinct partitions are aligned using majority voting, and decisions are made under the minimum class reconstruction error criterion.

2.2.4 Challenges of Facial Modeling

One of the main challenges of FRiVS is that facial models lose their representativeness over time because they are designed *a priori* design using a limited number of reference samples captured under semi- and uncontrolled conditions. Facial captures incorporate considerable variations because of the limited control over operational conditions in the scenes – changes in illumination, pose, facial expression, orientation, occlusion, etc. Furthermore, the physiology of enrolled individuals may change over time, either temporarily (e.g., hairstyle, cosmetics, glasses, etc.) or permanently (e.g., aging, surgery, etc). These factors result in facial models that are not representative of faces to be recognized. However, new information may emerge during operations to update of re-enrollment, and formerly collected data may eventually become obsolete in a changing environment. As described in Section 2.3, several adaptive biometric techniques have been proposed to update biometric models over time, and maintain or improve a high level of performance.

2.3 Adaptive Biometric Systems

The internal structure of biometric models dictates the most effective strategy for adaptation. In general, it involves (1) the *selection* of diversified, relevant reference samples to update a template gallery or an LTM of reference validation samples, and (2) the actual *update* of template galleries or classifier parameters using supervised or semi-supervised learning schemes.

2.3.1 Selection of Representative Samples

In this chapter, adaptive MCS are considered for FRiVS, where an ensemble of detectors (EoD with 2-class classifiers trained on target vs. non-target samples) is used to design the facial models of individuals of interest (De-la Torre *et al.*, 2012a). The level of informativeness of an input sample **a**, may be estimated using selection techniques based on the data itself, or using information retrieved from the ensemble. Examples of selection techniques used for FR

include editing algorithms such as the CNN, used to manage a gallery of templates in template matching systems (Freni *et al.*, 2008).

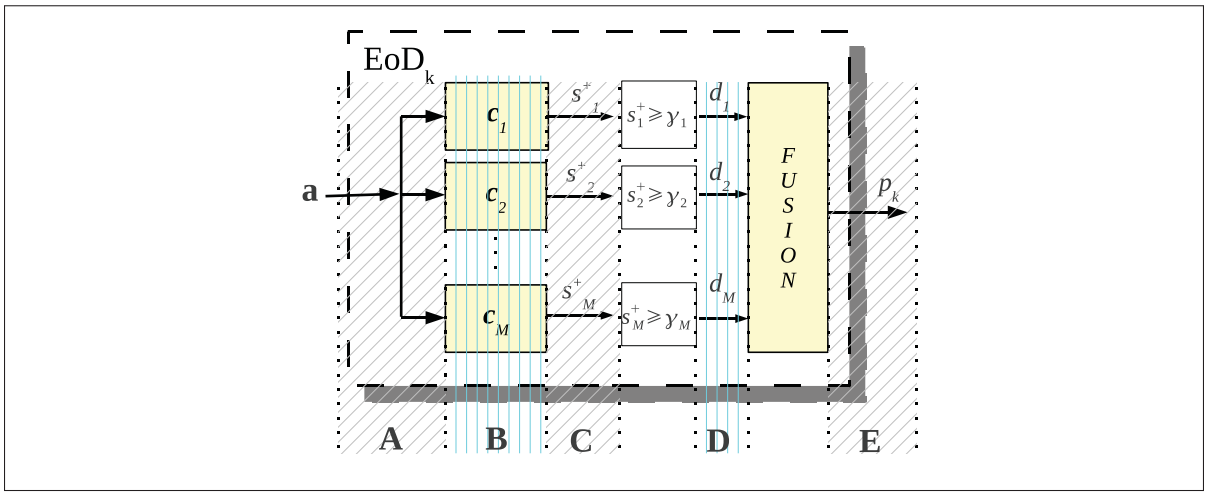


Figure 2.2 Ranking levels that are relevant for an ensemble of 1- or 2-class binary classifiers, e.g., for individual k

Fig. 2.2 presents the levels of selection that are relevant for ensembles of 1- or 2-class binary classifiers. The *input data level* **(A)** allows to use the dataset itself to filter out redundant samples. At this level, the estimation of the data distribution of samples is not required in the filtering process, which makes the methods at level **(A)** dependent only on the reference samples. Filtering methods here do not use a ranking, but rather, the geometric relationship between samples in feature space. At the *classifier level* **(B)**, the relevance measure of samples is retrieved from the internal response of the classifier to an input sample **a**. At the *classifier score level* **(C)**, the output scores $s_m^+(\mathbf{a})$ of the M classifiers in the ensemble are combined to produce a measure of relevance. When probabilistic classifiers are used as base classifiers, the relevance measure computation is based on the combined estimated posterior probability (classification scores s_m^+). At the *classifier decision level* **(D)**, the decisions $d_m(\mathbf{a})$ of the classifiers in the ensemble are combined. Voting strategies can be used to generate a relevance measure such as vote entropy. Finally, at the *ensemble decision level* **(E)**, the global output of the ensemble can be used as a measure of informativeness of the input sample.

Table 2.1 Sampling techniques for the selection of representative samples according to the five ranking levels from Fig. 2.2

Technique	**A**	**B**	**C**	**D**	**E**
Uncertainty sampling (from Active Learning)					
Less confident (Lewis and Catlett, 1994)					✓
Surprise (Liu *et al.*, 2010)					✓
Margin Sampling (Scheffer *et al.*, 2001)		✓			✓
Entropy Sampling (Shannon, 1948)					✓
Query by Committee (from Active Learning)					
Average surprise (Liu *et al.*, 2010)			✓		
Average Margin Sampling (Scheffer *et al.*, 2001)		✓	✓		
Vote Entropy (Dagan and Engelson, 1995)				✓	
Kullback-Leibler divergence (Kachites McCallum and Nigam, 1998)			✓		
Other measures inspired in diversity of ensembles					
Margin (voting) (Tang *et al.*, 2006)				✓	
Less confident (voting) (Lewis and Catlett, 1994)				✓	
Surprise (voting) (Liu *et al.*, 2010)				✓	
Resampling techniques (Guo *et al.*, 2008; Galar *et al.*, 2011)					
Condensed Nearest Neighbor rule (Hart, 1968)	✓				
Random Undersampling	✓				
SPIDER	✓				
One-Sided Selection	✓				
Wilson's Edited Nearest Neighbor rule	✓				
Neighborhood Cleaning Rule	✓				
Tomek links (Tomek, 1976)	✓				
Boosting weighting	✓				
Budget-sensitive, progressive-sampling	✓				

Table 2.1 presents sampling techniques from the literature according to the five ranking levels. Techniques that operate at level **A**, are suitable when the distribution of the new incoming data is unknown, e.g., before the samples are used in the design/update process. Using data dependent techniques to select reference samples avoids any bias produced by the knowledge already embedded in the system. At level **B**, information from the internal components of the classifiers are used to estimate the relevance of test samples. However, given that such information is incompatible from one classifier to another, such ranking techniques usually suffer from poor representativeness of the informativeness of a sample.

Levels **C** and **D** are independent of the classification algorithm used in the ensemble and in the combination strategy. The only constraint imposed at level **C** lies in the compatibility of scores produced by classifiers, a limitation that can be overcome by using normalization strategies. Alternatively, probabilistic base classifiers can be used, taking advantage of their output estimated posterior probabilities, and avoiding the need for normalization. Level **D** is also a good candidate for combining decisions from (crisp) classifiers; however, the resolution is limited by the number of classifiers in the ensemble. Finally, level **E** estimates the informativeness of an input sample using information from ensemble members and the fusion function. Crisp decision functions, such as the weighted majority voting or Boolean combination, provide a decision that can produce a binary relevance measure. Otherwise, it must be converted to a score in order to be used as a multiple-valued relevance measure (e.g. using the ROC space (Flach and Matsubara, 2008)). In that case, an extra validation set may be required, which is impractical in many real applications.

Given a set of positive target samples, and the availability of abundant non-target samples in the application (the CM and UM), the selection of a representative subset of representative training samples becomes essential for practical implementations. Level **A** in Fig. 2.2 provides a wide spectrum of techniques, in which different approaches allow for the selection of samples from distinct regions of the data distributions. For instance, the CNN finds the borderline samples, whereas using Tomek Links allows to remove both noisy and borderline samples from the set of data. On the other hand, one sided selection allows to remove noisy and borderline samples from the majority class by combining Tomek Links followed by CNN. Due to the complexity of the non-target distribution (e.g. it holds samples from all non-target individuals), non-target borderline samples are important for classifier training. These samples allow for a fine tuning of the decision frontier between classes. In this chapter, the CNN has been used to select borderline samples between target and non-target data distributions, providing more relevance to the samples closer to the overlapping area (Hart, 1968). In Section 2.4, a CNN-based strategy is proposed to consider representative samples from the target and non-target distributions, and especially those samples in their overlapping zone.

Different from uninformed selection (level **A**), an informed selection of validation samples considers the responses of the base classifiers in the ensemble, and takes advantage of the current state of knowledge of the classification system. From the rest of the (informed) ranking levels, level **B** is not considered because of the incompatibility of the internal information between classifiers. And level **E** is not considered given that the information is reduced to a single decision, and an extra validation set may be required to produce a multi-level ranking. After this reasoning, ranking measures from levels **C** and **D**) are chosen as best candidates. The graphs of the measures at these levels were analyzed (see I), and it can be seen that average margin sampling (AMS), Kullback-Leibler (KL) divergence and vote entropy (VE) present a peak in the overlapping region between target and non-target distributions. These samples in the overlapping region are of special interest for validation given that they provide a higher level of information. From the aforementioned measures, VE shows a lower resolution than KL and AMS, and the smoothness of the KL divergence curve shows a better representation of the overlapping area. Furthermore, the KL divergence takes advantage of the posterior probabilities estimated by the base classifiers, and allows to select the samples that provide the highest level of information, which appear in the overlap areas between classes, close to the decision boundaries. In this chapter, the KL divergence is employed to implement a strategy for assessing the relevance of reference samples in managing a fixed size memory of validation samples.

2.3.2 Update of Biometric Systems

In the literature, several approaches allow for supervised adaptation providing reliable results (De-la Torre *et al.*, 2012a; Connolly *et al.*, 2012; Tax and Duin, 2008), and yet obtaining labeled reference samples is costly or impractical. To overcome this difficulty, some *semi-supervised* methods have been introduced for automatic template updates (Roli and Marcialis, 2006; Franco *et al.*, 2010; Roli *et al.*, 2007, 2008; Okada *et al.*, 2001; Rattani *et al.*, 2008a, 2009b). This chapter focuses on the semi-supervised updating of biometric models. *Self-*

training and *co-updating* are two well-known algorithms for semi-supervised adaptation using template matching.

In *self-update* methods (Roli *et al.*, 2007), the biometric models are first designed storing samples from a labeled data set D_L in a template gallery $\mathscr{G}$. Prediction is possible by applying a decision threshold γ^d to the similarity score produced after template matching. Then, during operations, similarity scores are produced for the unlabeled samples, and those with a high degree of confidence (surpassing an updating threshold $\gamma^u \geq \gamma^d$), are integrated to the gallery $\mathscr{G}$, thereby updating the corresponding biometric models. The notion of "high degree of confidence" is subjective, and depends on both the matching algorithm and the application domain, but an update threshold higher or equal than the prediction threshold is commonly used. This procedure is detailed in Algorithm 2.1.

Algorithm 2.1: Self-update algorithm to adapt a gallery for template matching

Input :
  $\mathscr{G} = \{t_1, ..., t_N\}$ // Gallery with initial templates
  $D = \{d_1, ..., d_L\}$ // Unlabeled adaptation set
Output :
  $\mathscr{G}' = \{t_1, ..., t_N, ..., t_M\}, M \geq N$ // Updated template gallery
Estimate threshold $\gamma^u \geq \gamma^d$ *for the templates in* $\mathscr{G}$
$\mathscr{G}' \Leftarrow \mathscr{G}$ // Initialize with $\mathscr{G}$
// For all samples $d_l \in D$
for $l = 1, ..., L$ **do**
  // For all templates in the gallery $t_l \in \mathscr{G}$
  for $n = 1, ..., N$ **do**
    $s_{n,l} \Leftarrow similarity\_measure(d_l, t_n)$ // Compute score against all samples in $\mathscr{G}$
  $s_l \Leftarrow \max\{s_{n,l} : n = 1, ..., N\}$
  if $s_l > \gamma^u$ **then**
    $\mathscr{G}' \Leftarrow \mathscr{G}' \cup d_l$ // Include the sample surpassing γ^u in the new data set

Co-update is a semi-supervised learning strategy adapted for use with two diversified matchers with galleries specialized on distinct biometric traits, which are designed to improve performance mutually (Roli *et al.*, 2007). For example, in (Roli *et al.*, 2007), authors propose the use of fingerprints and the face, using co-training for semi-supervised updates of the facial and fingerprint models. Algorithm 2.2 presents the co-training algorithm. The procedure starts with

the design of the two matchers with the labeled templates in galleries $\mathscr{G}_1$ and $\mathscr{G}_2$, and selecting ad-hoc the thresholds for decision (γ_1^d and γ_2^d) and update (γ_1^u and γ_2^u). Once the unlabeled sets D_1 and D_2 are collected, both matchers are used to label the samples, and those with high degrees of confidence (at least in one of the matchers) are added to the updated galleries $\mathscr{G}_1'$ and $\mathscr{G}_2'$. Also the decision and update thresholds are be updated over time in accordance with the newly acquired data. A potential advantage of the co-update algorithm is that it can retrieve update samples that are not typical of the distribution of target data from a single trait, allowing adaptation to diverse, possibly abrupt changes.

The advantages of adapting a biometric system using operational data carries an inherent risk. There exists a trade-off between the false updates and false rejections that affect of performance. A conservative threshold (or other parameters in the biometric model) may allow a system without false updates, but also a system that is never adapted to changes in the environment. Conversely, a less conservative threshold may contribute to increase in the number of false updates and the inherent deterioration of biometric models. Following this reasoning, we can easily see that a good selection of adaptation criteria (decision threshold) is crucial in the design of the system.

Other semi-supervised approaches take advantage of neural or statistical classifiers in the construction of biometric models. For instance, in (Okada *et al.*, 2001), a view representation that combines facial and torso-color histograms was used with bunch graph matching for adaptive person recognition. The system is capable of updating existing biometric models and to automatically enroll unknown individuals based on a double thresholding strategy. Update was performed on operational video streams that provide high sequence-to-entry similarity, measure of confidence. The sequence-to-entry similarity is the average of maximum frame-to-entry similarity values, which in turn was defined as the maximum similarity value over all facial representations in a database entry (Okada *et al.*, 2001). Bayesian networks were also used to recognize facial expression and detect faces using a stochastic structure search algorithm (Cohen *et al.*, 2004). This approach combined labeled and unlabeled samples to train the Bayesian networks, and seek for the Bayesian network structure that provided the minimum

Algorithm 2.2: Co-update algorithm to adapt a gallery for template matching

Input :
$\mathscr{G}_1 = \{t_1^1, ..., t_{N_1}^1\}$
and $\mathscr{G}_2 = \{t_1^2, ..., t_{N_2}^2\}$ // Galleries with initial templates
$D_1 = \{d_{1,1}, ..., d_{L,1}\}$
and $D_2 = \{d_{1,2}, ..., d_{L,2}\}$ // Unlabeled adaptation sets, $d_{l,1}$ corresponds to $d_{l,2}$
Output :
$\mathscr{G}_1' = \{t_1^1, ..., t_{N_1}^1, ..., t_{M_1}^1\}$,
$M_1 \geq N_1$ // Updated galleries for both modalities
$\mathscr{G}_2' = \{t_1^2, ..., t_{N_2}^2, ..., t_{M_2}^2\}$, $M_2 \geq N_2$
Estimate thresholds $\gamma_1^u \geq \gamma_1^l$ *and* $\gamma_2^u \geq \gamma_2^l$ *for the* $\mathscr{G}_1$ *and* $\mathscr{G}_2$ *respectively*
// For each gallery G_i, $i = 1,2$
for $i = 1,2$ **do**
  $\mathscr{G}_i' \Leftarrow \mathscr{G}_i$ // Initialize with templates in the gallery i
  // For all samples $d_{l,i} \in D_i$
  for $l = 1, ..., L$ **do**
    // For all templates in the gallery $t_{n,i} \in \mathscr{G}_i$
    for $t_{n,i} \in \mathscr{G}_i$, $n = 1, ..., N_i$ **do**
      $s_{n,l,i} \Leftarrow similarity\_measure(d_{l,i}, t_{n,i})$ // Compute score for all $d_n \in D_i$
    $s_{l,i} \Leftarrow \max\{s_{n,l,i} : n = 1, ..., N_i\}$
    if $s_n^i > \gamma_i^u$ **then**
      $j \Leftarrow \mod(i+1,2)+1$ // Samples added to the complementary gallery
      $\mathscr{G}_j' \Leftarrow \mathscr{G}_j' \cup d_{l,j}$

probability of error, using maximum likelihood estimation. SVMs with locality preserving projections have also been combined to update facial models, by incorporating information from operational ROIs taken from video (Lu *et al.*, 2010). The algorithm first builds a data model of a video sequence, and then uses semi-supervised locality preserving projections to assemble a graph with the geometrical structure of the feature space of faces.

MCSs have also been used in conjunction with the co-training and self-training. In (Didaci and Roli, 2006), for instance, an ensemble of five classifiers was trained with two different diversity generation techniques (bootstrap and the training of different classifiers). These techniques are based on a re-training schema for biometric model updates, and improve accuracy by 18% using the product rule for combination. Another modification of the co-training algorithm for MCS was proposed for updating only unlabeled samples that produced high confidence (El Gayar *et al.*, 2006). The five patterns with highest probability of belonging to the specific

person, were selected as the most confident. This system was tested with 3 non-homogeneous classifiers in the ensemble, and provided the highest performance with a voting combination scheme. Finally, a semi-supervised classification schema based on random subspace dimensionality reduction was proposed for graph-based semi-supervised learning. In this approach, a kNN graph is built in each processed random subspace, and semi-supervised classifiers are trained on the resulting graphs, using majority voting rule for combination (Yu *et al.*, 2012).

MCSs for semi-supervised learning in the literature have provided improved accuracy, and show the utility of unlabeled samples. In this chapter, an adaptive MCS is proposed for spatio-temporal FR, that allows for semi-supervised learning from facial trajectories defined by the face tracker. It exploits the two thresholds (γ^d and γ^u) from the self-update algorithm, and the quality of tracking as a second source of confidence, characteristic borrowed from the co-update algorithm. The tracking quality allows to regroup facial regions from the same individual, and the accumulation of the predictions from the user-specific ensembles over time allow for high confident decisions.

2.3.3 Adaptive Face Recognition

In the literature, adaptive FR systems have traditionally incorporated new training data to update the selection of templates from a facial database, using clustering and editing techniques. Processing thus allows an improved representation of intra-class variations to be obtained using a sole template. These systems were proposed to improve facial models considering the intra-class variations from input samples (Roli *et al.*, 2008).

Recent work on the *supervised update* of facial models includes an FR system formed from an adaptive MCS. A DPSO based incremental learning strategy has been proposed for video-based access control. It allows the evolution of an ensemble of heterogeneous multi-class classifiers from new data, using an LTM to store validation samples for fitness estimation and to stop training epochs. This approach reduces the effect of knowledge corruption (Connolly *et al.*, 2012). Another adaptive MCS for designing and updating facial models is composed of an

EoD per individual, an LTM and a dynamic optimization based training module. When a new data block becomes available, a diversified pool of ARTMAP neural networks is generated by a DPSO based learning strategy. The combination function is updated using Boolean combination (BC) (De-la Torre *et al.*, 2012a). Learn++ is another ensemble-based incremental learning technique that has been tested on FR problems (Polikar *et al.*, 2001). It performs supervised incremental learning by training and integrating a new batch of weak classifiers to the ensemble when new reference samples become available. These weak classifiers are generated using a bagging strategy inspired in the AdaBoost algorithm.

Semi-supervised approaches for facial model update are generally based on the classification similarity. For instance, in (Roli and Marcialis, 2006), semi-supervised learning has been applied to FR with self-training, using an Euclidean distance-based measure of similarity. In each iteration, the PCA-based feature space is updated with the newly acquired soft-labeled samples. In (Hewitt and Belongie, 2006), the authors propose a method for combining tracking and recognition to build a facial model based on co-training. This method is used to label face samples and thus to build a learning dataset for each user. Their initial facial model consists of a single manually selected frontal face picture, and the extraction of new face samples is done off-line. In order to identify informative training samples, they replace the second classifier with a tracker. An extension to the self-update algorithm named the Graph Mincut (Rattani *et al.*, 2008a), has been proposed to update templates. This approach analyzes the underlying structure of operational data, and a pair-wise similarity measure between operational data and existing templates is used to draw a graph that relates these samples.

A representative example that exploits not only the classification similarity, but also video information, is presented in (Franco *et al.*, 2010). The authors propose an update strategy called incremental template update. It is based on the similarity between input samples and gallery templates. It exploits the frequency of detection on the complete sequences for the individuals in front of the camera, and combines this frequency with the coordinates of the detection within the last frame in the sequences.

2.4 A Self-Updating System for Face Recognition in Video Surveillance

In this chapter, an adaptive MCS is proposed for spatio-temporal FRiVS that allows for partially-supervised learning from facial trajectories. As shown in Fig. 2.3, the proposed system is comprised of a segmentation module for face detection, a face tracker, a modular classification system with one EoD per individual of interest, a decision fusion system, a design/update system, and a sampling selection system.

During operations, informations from a tracker and modular classifiers (user-specific EoDs) are integrated at a decision fusion level for enhanced video-to-video FR. A *highly confident* trajectory[2] T is associated with an individual of interest k when the number of accumulated positive predictions of a EoD over a fixed-size window of ROIs surpasses a predefined *detection* threshold (γ_k^d).

The MCS allows for self-update of facial models over time, based on diverse ROIs captured within trajectories. When an individual of interest k is detected by the system within a high quality trajectory T, and the number of positive predictions surpasses a second higher *updating* threshold, $\gamma_k^u \geq \gamma_k^d$, all the corresponding facial ROIs are combined (as target samples) with selected non-target samples from the CM and UM to produce a labeled training data set D to update a facial model. User-specific EoDs are updated using a *learn-and-combine* strategy, thereby avoiding knowledge corruption (De-la Torre *et al.*, 2012a). A new pool of detectors (2-class classifiers) is generated with D, and combined with previously learned detectors to adapt the EoD. For an accurate estimation of a fusion function and selection of an operations point, the LTM stores and updates a representative set of validation samples. Finally, a strategy based on Kullback-Leibler divergence is employed to rank and store only the most representative facial samples from the LTM. It combines ROI matching scores of user-specific ensembles within high quality facial trajectories captured with a tracker, for efficient self-updating of facial models over time. The set of ROIs associated with trajectories provide diversity for robust EoDs design.

[2]The notation T_k is reserved for trajectories assigned to an individual of interest k, for a design-update phase, e.g. labeled trajectories, whereas T is used for unlabeled operational trajectories.

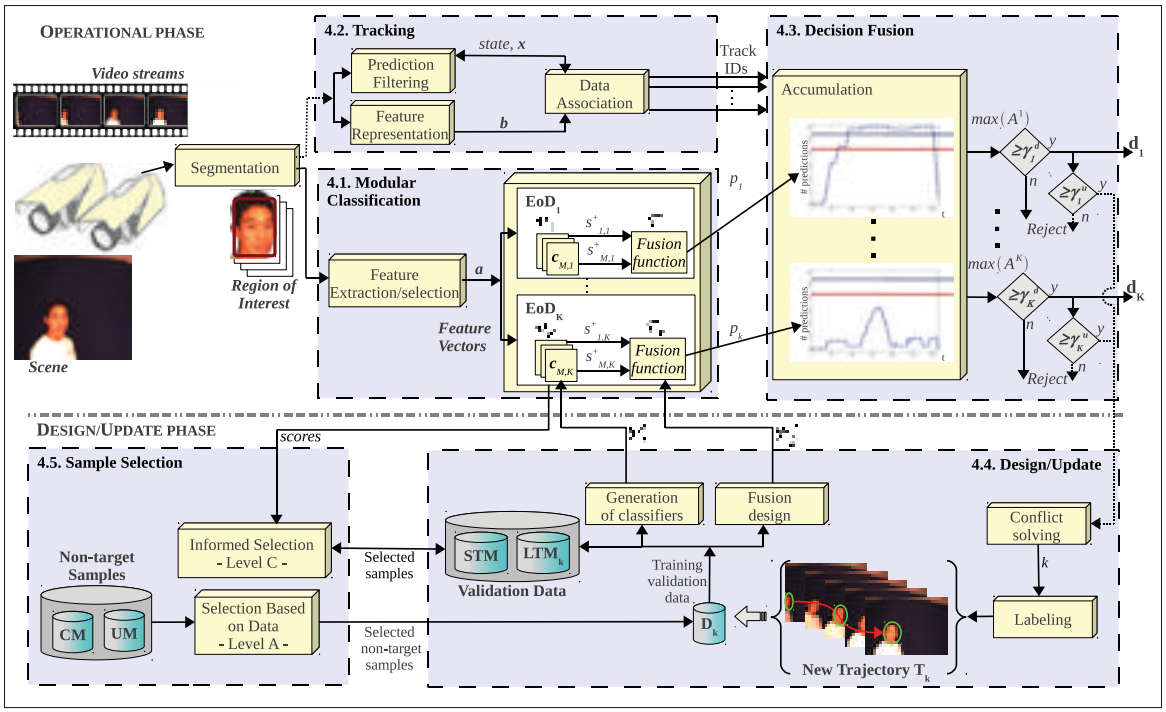


Figure 2.3 Block diagram of the proposed self-updating system for spatio-temporal FR in video surveillance

2.4.1 Modular Classification System

A modular classification architecture is proposed in this chapter. Individual-specific EoD allow for enhanced classification accuracy when only a limited number of training samples is available for system design (Pagano *et al.*, 2012). Accordingly, each EoD estimates discriminant bounds between the target (individuals of interest) and non-target (the rest of the world) classes. Each ensemble EoD_k is comprised of a pool of 2-class classifiers $\mathscr{P}_k = \{c_{1,k}, ..., c_{M,k}\}$, and a fusion function $\mathscr{F}_k$ that is designed using a validation set D_k^c, for $k \in \{1, ..., K\}$.

During operations, each ensemble member $c_{m,k}$ produces an output score $s_{m,k}^+(\mathbf{a})$ for a given feature vector $\mathbf{a}$ corresponding to an input ROI. The scores are then combined using $\mathscr{F}_k$. Each individual-specific EoD_k produces an output prediction $p_k(\mathbf{a})$. Positive predictions are then accumulated over time in the decision fusion system to produce a composed decision (see Fig. 2.3).

The fusion function $\mathscr{F}_k$ holds a set of operations points. Each point is comprised of classifier specific thresholds and combination functions (e.g. a Boolean combination or voting scheme). Depending on the strategy used for the estimation of the fusion function, a subset of the classifiers in the pool $\mathscr{P}_k$ is selected to maximize performance. The evaluation of the operations points on a selection set D_k^s allow to select a specific operations point in the ROC space, given a predefined acceptable fpr. Given that the system seeks to maximize the tpr under a constraint of the amount of false positives, the convex hull is selected in order to consider only the points with highest tpr. If there is no operations point for a specific fpr, a virtual classifier is produced by interpolating the closest adjacent operating points (Fawcett, 2006).

Finally, the self-update is achieved by using adaptive EoDs, each one is capable of supervised incremental learning. A *learn-and-combine* strategy is employed to maintain performance even after several adaptations, yet avoid knowledge corruption associated with many incremental learning classifiers (De-la Torre *et al.*, 2012a).

2.4.2 Tracking System

As shown in Fig. 2.4, the face tracker initializes a new trajectory with the first facial ROI captured by the segmentation system in a different area of the scene. As the tracker follows the facial region through the scene, the segmentation system captures high quality facial ROIs for some of the frames, allowing to produce a trajectory (a trajectory T is defined over consecutive frames). Note that the segmentation module does not retrieve a facial region from all frames. The diverse set of facial ROIs belongs to the same individual is defined by the tracker. When the tracking quality Q_T falls under a (manually) pre-defined overall quality threshold ($Q_T < \gamma^T$), its trajectory is dropped.

2.4.3 Decision Fusion System

The adaptive MCS detects the presence of individuals of interest based on the number of positive EoD_k predictions over trajectories. Given a high quality trajectory T, each EoD_k generates

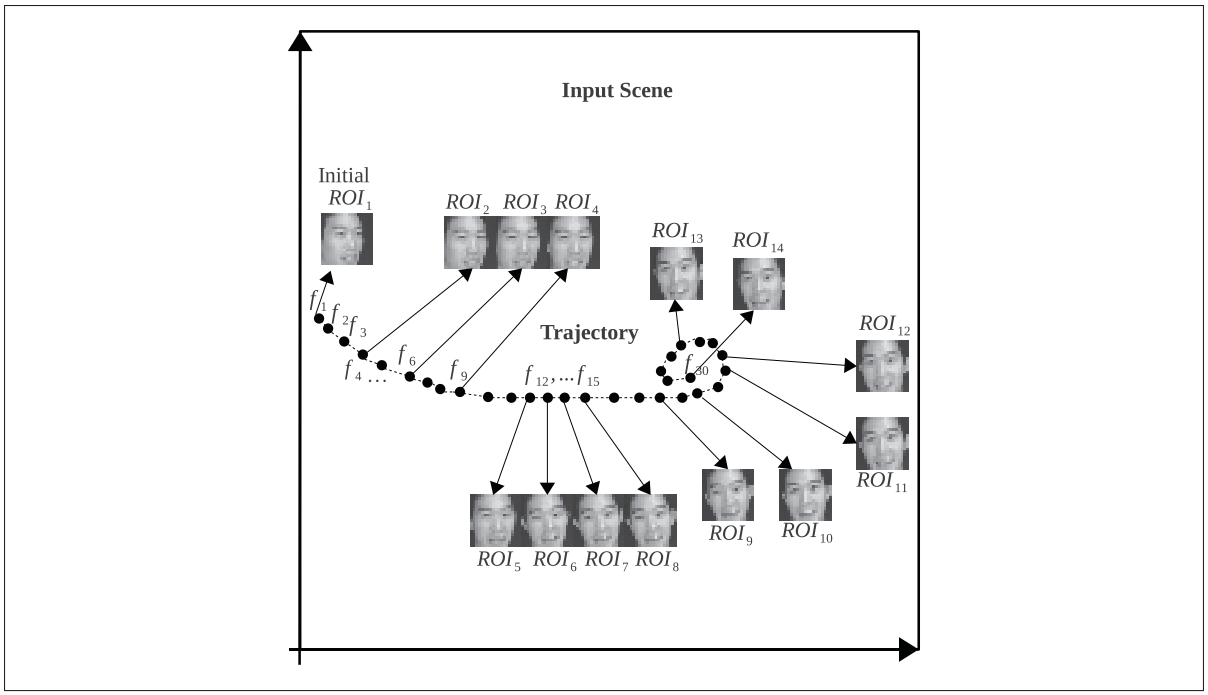


Figure 2.4 Illustration of the trajectory formation process within 30 frames of a FIA video. The tracker is initialized with ROI_1 and follows the face of an individual (person with ID 2), through the scene (capture session 1). f_i represents the position of the face in the camera view for frame i. The ROIs in the trajectory are produced by segmentation at $f_1, f_4, f_6, ..., f_30$, and the track is dropped at f_30. The trajectory is $T = \{ROI_1, ROI_2, ..., ROI_{14}\}$

a prediction $p_k(\mathbf{a}_n)$ for each sample $\mathbf{a}_n$ associated with a ROI in the trajectory. Output predictions from EoD_k over the ROI samples of a trajectory T, at the selected operations point, are defined by the set $\mathbf{P}_k = \{p_k(\mathbf{a}_1), ..., p_k(\mathbf{a}_N)\}$, associated with each input ROI sample $\mathbf{a}_n$. Negative predictions set $p_k(\mathbf{a}_n) = 0$, and positive ones set $p_k(\mathbf{a}_n) = 1$. The decision fusion system accumulates the number of positive predictions A_k of each EoD_k on fixed size window W according to:

$$A_k = \sum_{i=0}^{W-1} p_k(\mathbf{a}_{(W-i)}) \quad \in [0, W] \tag{2.1}$$

For instance, a window of size $W = 30$ accumulates the last 30 positive predictions from the same trajectory. Each EoD_k accumulates a sequence of positive predictions that range from 0

(EoD_k made only negative predictions for W), to a maximum of W (EoD_k made only positive predictions for the last W ROIs).

Based on these accumulations A_k, for $k = 1, ..., K$, the system produces decisions. If A_k surpasses threshold γ_k^d, the system detects the presence of individual k and alerts the operator. Furthermore, if A_k surpasses the update threshold γ_k^u, the trajectory is suitable for self-updating of the corresponding EoD_k. Given the negative effects on performance caused by false updates, threshold γ_k^u is greater or equal to γ_k^d.

For each EoD_k, the detection threshold γ_k^d is estimated using a validation set composed of one positive and several negative trajectories. In this way, a single target trajectory is required for design and update of the facial model. An accumulation curve is computed for each trajectory in the validation dataset. The *higher negative envelope* (*hne*) is defined as the curve formed from the highest A_k values of the negative accumulation curves. The detection threshold for EoD_k is computed as the maximum value in the *hne* plus the maximum difference between the *hne* and the *positive accumulation curve* (*pac*) for the corresponding individual k:

$$\gamma_k^d = \max\{hne(f_i) : i = 1, ..., |T_k|\} + \left(\frac{\max\{pac(f_i) - hne(f_i) : i = 1, ..., |T_k|\}}{2} \right) \tag{2.2}$$

where f_i is the frame number i in the trajectory. By considering the presentation order of the target (positive) and non-target (negative) facial regions, the time information is included in the threshold estimation for specific facial models. The adaptation threshold γ_k^u is set to a value equal to or greater than γ_k^d:

$$\gamma_k^u = \gamma_k^d + \Gamma_k \tag{2.3}$$

where Γ_k is a user-defined real value between 0 and $(W - \gamma_k^d)$. Fig. 2.5 illustrates the measures used in the threshold estimation strategy, presenting the *pac* and the *hne*. The reliability of γ_k^d and γ_k^u estimates grows with the number of non-target trajectories present in the validation set.

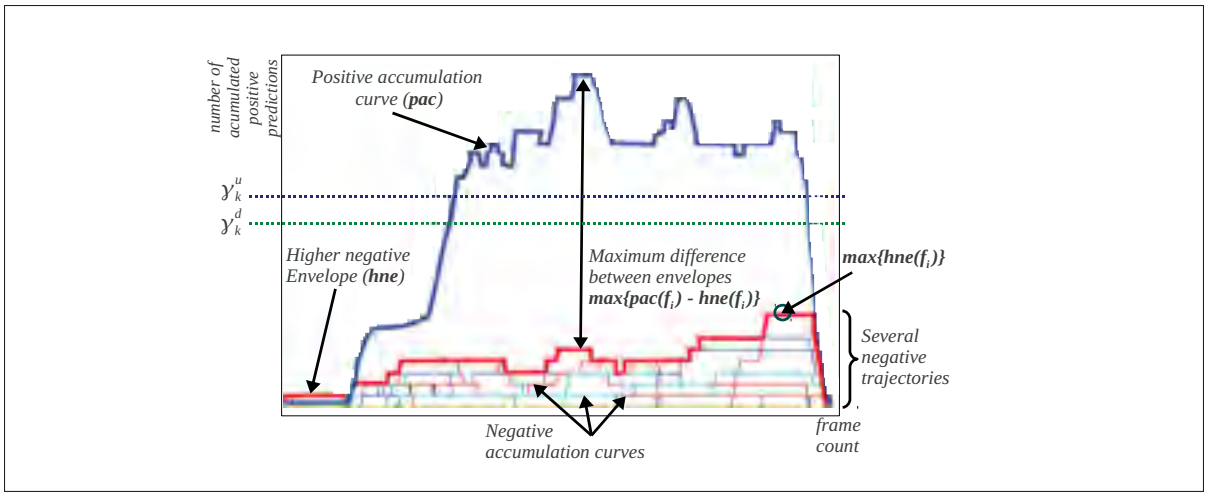


Figure 2.5 Detection and update threshold estimation on validation trajectories at the decision level

When the accumulative curve corresponding to an operational trajectory T surpasses the detection threshold γ_k^d for one or more EoDs, the system outputs the corresponding decision signals. The output to the decision support system lists all individuals of interest that are detected in the scene.

2.4.4 Design/Update System

Given a trajectory T, if the number of accumulated positive predictions from the EoD_k surpasses the update threshold, $A_k \geq \gamma_k^u$, the *design/update* system assigns the corresponding label to the trajectory. If conflict occurs (two or more EoDs detect the same trajectory as suitable for update), the EoD_k with highest A_k value is selected. If two or more trajectories present the same A_k value, the system is prevented from updating, and these conflicting trajectories are stored for further analysis by a human expert.

Once the trajectory has been successfully tested for conflicts, the system assigns the label k to all the patterns corresponding to the facial ROIs of the trajectory T, and it becomes a labeled trajectory T_k. An advantage of the proposed system is the incorporation of diversified information into facial models of detected individual. Self-updating provides EoDs with a

greater diversity of samples captured under various conditions (pose, lighting, etc). These samples allow for a more accurate definition of the boundaries between target and non-target individuals in accordance with the most recent facial samples.

When a new trajectory T_k is detected and labeled for update, it is divided into three subsets in order to follow a *learn-and-combine* strategy. A CNN based selection algorithm allows to retrieve borderline and distinctive samples from the negative distribution, by selecting negative samples from the CM and UM (see Section 2.4.5). The CM database is comprised of a set of trajectories from the individuals of interest, excluding individual k; and the UM database is comprised of trajectories from other non-target individuals that represent the rest of the world, e.g. random individuals that appear frequently in the scene. The subset D^t is used for training[3], D^e for validation on the number of training epochs, and D_k^f for optimization of classifier hyperparameters. Then, some ensemble generation strategy (e.g. random subspace methods, boosting and bagging, (Kuncheva, 2004)) allows to generate a diversified pool of classifiers, and add them to the previous pool $\mathscr{P}_k$. The samples from the validation sets (D^e and D^f) are then mixed with samples from the LTM_k[4], stored to a short term memory (STM_k), randomized and divided into two subsets (D^c and D^s). The classifiers from the pool $\mathscr{P}_k$ and the fusion function $\mathscr{F}_k$ are selected and combined using D^c, and the operations point is selected using D^s. The process is repeated for all the EoDs. In summary, each EoD_k is updated with new ROIs from a trajectory T_k by generating new base classifiers, adding these to a pool $\mathscr{P}_k$, and updating the fusion function according to the old and new validation samples.

If the size of the LTM_k for EoD_k is λ_k, the size of the STM_k is chosen to be $2\lambda_k$ in order to store enough new and old validation samples. This follows the assumption that old (from LTM_k) and new samples are equally relevant. Then, the validation samples in the STM_k are ranked according to Eq. 2.4 (see Section 2.4.5), and the λ_k samples with the highest values are stored in the LTM_k.

[3]For simplicity of notation, the k has been omitted from all design data blocks, e.g. $D_k^t \equiv D^t$.

[4]Note that the LTM_k is initially empty, and filled with positive and negative samples after the initial design.

Algorithm 2.3: Design and update of a user-specific ensemble of detectors, EoD_k

Input : T_k, $EoD_k = \{\mathscr{P}_k, \mathscr{F}_k\}$, LTM_k, UM, CM
Output : EoD'_k, LTM'_k // Updated $EoD'_k = \{\mathscr{P}'_k, \mathscr{F}'_k\}$ and LTM'_k

Divide T_k in D^t, D^e, D^f evenly // T_k keeps only positive samples

$D^t \Leftarrow CNN\_NEG\_SEL(D^t, UM, CM)$ // Form 2-class data sets with target (+) vs.
$D^e \Leftarrow CNN\_NEG\_SEL(D^e, UM, CM)$ // non-target (-) samples (see Algorighm 2.4)
$D^f \Leftarrow CNN\_NEG\_SEL(D^f, UM, CM)$
$P'_k \Leftarrow \{c'_{1,k}, ..., c'_{M,k}\}$ // Generate a pool $\mathscr{P}'_k$ using D^t, D^e and D^f

$\mathscr{P}_k \Leftarrow \mathscr{P}'_k \cup \mathscr{P}_k$ // Combine old and new classifiers in the pool

$STM_k \Leftarrow D^e \cup D^f \cup LTM_k$ // Store old and new validation samples in STM_k
Divide STM_k in D^c and D^s evenly
$\mathscr{F}'_k \Leftarrow FUSION(D^c, D^s, fpr)$ // Estimate fusion function given a predefined fpr
$EoD'_k \Leftarrow \{\mathscr{P}'_k, \mathscr{F}'_k\}$ // Updated selection of classifiers and fusion function
$LTM'_k \Leftarrow KL\_SEL(STM_k, \lambda_k)$ // Use KL to replace samples in LTM_k with most
// informative in STM_k

2.4.5 Sample Selection

Sample Selection for Training. Positive samples from the aforementioned design/update trajectory T_k are coupled with negatives from the CM and UM to form the learning set D. Negative samples from the CM and UM are stored in a single global fixed size memory capable of storing recent facial captures from non-target individuals. The size of this memory should be determined according to system requirements, but it should be large enough to store trajectories from several non-target individuals. In practice, the UM can be regularly updated with trajectories from random or selected individuals (e.g. employees or frequent clients), and the CM is updated every time the system receives update trajectories. The CNN subsampling strategy (Hart, 1968) is employed to reduce the bias of training 2-class classifiers with imbalanced data sets (limited positive vs. abundant negative samples). This method selects those samples from

both classes that lie on the area of overlap or are difficult to classify (outliers). Nevertheless, these samples are complemented with distinctive samples from the underlying distributions. Distinctive samples are selected by storing all available positive references as well as a uniform sampling of negative ones from UM and CM after CNN selection. This CNN negative selection strategy resembles one sided selection in the application of CNN selection, however the CNN negative selection does not discard borderline samples, and includes distinctive samples through random selection. This permits the update of the ensemble considering not only the most relevant past and present samples close to decision bounds, but also typical samples distinctive of the most recent states of distributions of data.

The CNN negative selection strategy is detailed in Algorithm 2.4. When a trajectory T_k is provided to the system for training/update, the corresponding ROIs are used to build dataset of positive samples D^+, and a set D^- is formed with samples from the UM and CM. The CNN algorithm is then applied to $D^+ \cup D^-$ to select a consistent subset for design of the binary base classifiers. The resulting dataset D comprises three parts of equal size: (1) the complete set of positives D^+, (2) the negative samples selected by CNN (close to the decision boundaries) $D^{-'}_{cnn}$, and (3) a uniform random selection of non-borderline negatives D^-_d. In this way, D contains all target samples and twice more non-target samples. Algorithm 2.4 makes no assumptions concerning the probability distribution of the positive and negative samples, and permits an unbiased selection of negative samples, based solely on the distribution of the new samples.

Management of LTM_k. Level **C** ranking measures (see Section 2.3.1) permit the selection of samples from the LTM_k that are difficult to classify by the ensemble members (in Fig. 2.2). These samples are distinctive of the decision bound between the target and non-target classes, as estimated with the base classifiers in the EoD. The disagreement of base classifiers on a determined validation sample is proportional to its difficulty, give a degree of information for border specification when the fusion function is estimated. This is also valid for the accurate selection of operations points. Among ranking measures available in the literature, the Kullback-Leibler divergence produces a continuous measure of the disagreement between the ensemble members that covers the overlapping area between class distributions (see analysis in

Algorithm 2.4: *CNN_NEG_SEL.* Select negative samples to design the system

Input : D^+, UM // Positive and negative samples from UM and CM data bases

Output : D // Design dataset with all positive and selected negative samples

$D^- \Leftarrow UM \cup CM$ // Consider all negative samples from UM and CM

$[D^+_{cnn}, D^-_{cnn}] \Leftarrow CNN(D^+, D^-)$ // Samples selected by CNN

$np \Leftarrow |D^+|$ // Number of positive samples

$D^{-'}_{cnn} \Leftarrow RAND\_SEL(D^-_{cnn}, np)$ // Select np negatives from D^-_{cnn} belonging to UM and CM evenly

$D^-_d \Leftarrow RAND\_SEL(D^-, np)$ // Select np distinctive negatives from D^-, not selected by CNN

$D \Leftarrow D^+ \cup D^{-'}_{cnn} \cup D^-_d$

I). Accordingly, the KL divergence permits the exploitation of the knowledge from base classifiers to select the validation samples that provide the highest level of information. Even more, its continuous ranking values permit the discrimination between two samples that appear very close to each other in the feature space. The KL divergence of an input sample **a** is computed using:

$$KL(\mathbf{a}) = \frac{1}{M} \sum_{m=1}^{M} \left(\sum_{i \in \Omega} s^i_m(\mathbf{a}) \log \frac{s^i_m(\mathbf{a})}{\hat{P}^i_{EoD_k}(\mathbf{a})} \right) \tag{2.4}$$

where M is the number of classifiers in the ensemble EoD_k, and $\hat{P}^i_{EoD_k}(\mathbf{a})$ given by (2.5) is the consensus probability that the class $i \in \Omega$ is the correct label for sample **a**, given the scores $s^i_n(\mathbf{a})$ produced by the base classifiers:

$$\hat{P}^i_{EoD}(\mathbf{a}) = \frac{1}{M} \sum_{n=1}^{M} s^i_n(\mathbf{a}) \tag{2.5}$$

The value of KL divergence is proportional to the level of information provided by a sample **a**. The most informative samples present the largest average difference between scores of any single committee member and the consensus.

Algorithm 2.5 details the selection process that considers all the validation samples in the STM_k. Given an EoD_k, the *KL_SEL* algorithm selects the λ_k most challenging samples from the validation set, providing those samples lying on the overlapping area according to the agreement of the ensemble members. When a validation dataset D is presented to the algorithm, all samples are ranked according to the KL divergence using the scores produced by all the base classifiers in the pool $\mathscr{P}_k$. The λ_k highest ranked samples are retained, while the less informative ones are discarded, maintaining the proportion of target and non-target samples. Thus, the ranking method is based on past and present information on samples that are difficult to classify, according to older and newer classifiers.

Algorithm 2.5: Subsampling using the KL divergence, $KL\_SEL(input = \{D, s_k(a_i), \lambda_k\}, output = \{Dr\})$

Input : $D, s_k(a_i), \lambda_k$ // Data block, scores $s_k(a_i)$, $a_i \in D$ produced by EoD_k and size of the LTM_k

Output : Dr // Data block with λ_k representative samples from D

// For each sample in the data block

for $a_i \in D$ **do**

  $relevance_i = KL(s_k(a_i))$ // Compute the KL divergence according to Eq. 2.4

$D \Leftarrow SORT(D, relevance, dec)$ // Sort D in decreasing order, according to $relevance_i$

$Dr^+ \Leftarrow FIRST\_POSITIVES(D, \lceil \frac{\lambda_k}{2} \rceil)$ // Positive samples with highest KL divergence

$Dr^- \Leftarrow FIRST\_NEGATIVES(D, \lceil \frac{\lambda_k}{2} \rceil)$ // Negatives with highest KL divergence

$Dr \Leftarrow Dr^+ \cup Dr^-$

2.5 Experimental Methodology

Some methodologies for performance evaluation of adaptive biometric systems divide the design-update data into subsets, and use a same independent test set to show the evolution of performance (Singh *et al.*, 2010; Roli and Marcialis, 2006; Franco *et al.*, 2010; Liu and Cheng, 2003). Others divide the unlabeled data set into subsets, and progressively update on

a subset while testing on the next subset (Roli *et al.*, 2007; Rattani *et al.*, 2008a). This last approach is followed in this chapter. The main task under evaluation is detecting the presence of individuals of interest in semi-constrained environments, and the experimental protocol was designed to study the evolution of system performance in a changing classification environment. The adaptive MCS is first trained using design trajectories from an enrollment session (D), then the updating process was performed on three different capture sessions D_t, $t = 1...3$ in a video-to-video recognition scheme. The system is adapted after the presentation of each session D_t with the trajectories detected as positives, and the performance is evaluated using ROC and PROC spaces after the presentation of a different capture session D_{t+1}.

2.5.1 Video Surveillance Database

The proposed system was characterized in a video surveillance scenario using the Carnegie Mellon University Face in Action (FIA) database (Goh *et al.*, 2005). The FIA database contains 20-second videos that capture the faces of 180 participants that simulate a passport checking scenario. Capture speed is fixed to 30 frames per second, with a resolution of 640×480 pixels. An array of 6 cameras was positioned at the face level to capture the scene. However only the 2 frontal cameras are considered here. They are positioned at 0^o (frontal) and $\pm 72.6^o$ angle with respect to the individual. Three of the cameras were set at a zoomed focal-length (8-mm), resulting in face areas over 300×300 pixels. The other three cameras were set at an unzoomed focal length (4-mm), resulting in face areas over 100×100 pixels. Data was captured in three sessions separated by a three-month interval for each individual. Facial regions of interest (ROIs) were detected in videos using the Viola-Jones algorithm (Viola and Jones, 2004). Visual tracking was also applied on video sequences, initializing the Continuously Adaptive Mean Shift (CAMSHIFT) (Bradski, 1998) with the first face detected. All images were scaled to 70×70 pixels, which is the maximum resolution of the smallest face detected by the Viola-Jones algorithm. The Multi Scale Local Binary Patterns (MS-LBP) (Ojala *et al.*, 2002) feature extractor was used with three block sizes (3x3, 5x5 and 9x9), in conjunction to pixel-intensities

features. These features were stacked, and the 32 principal characteristics (PCA) were selected to form the feature vectors.

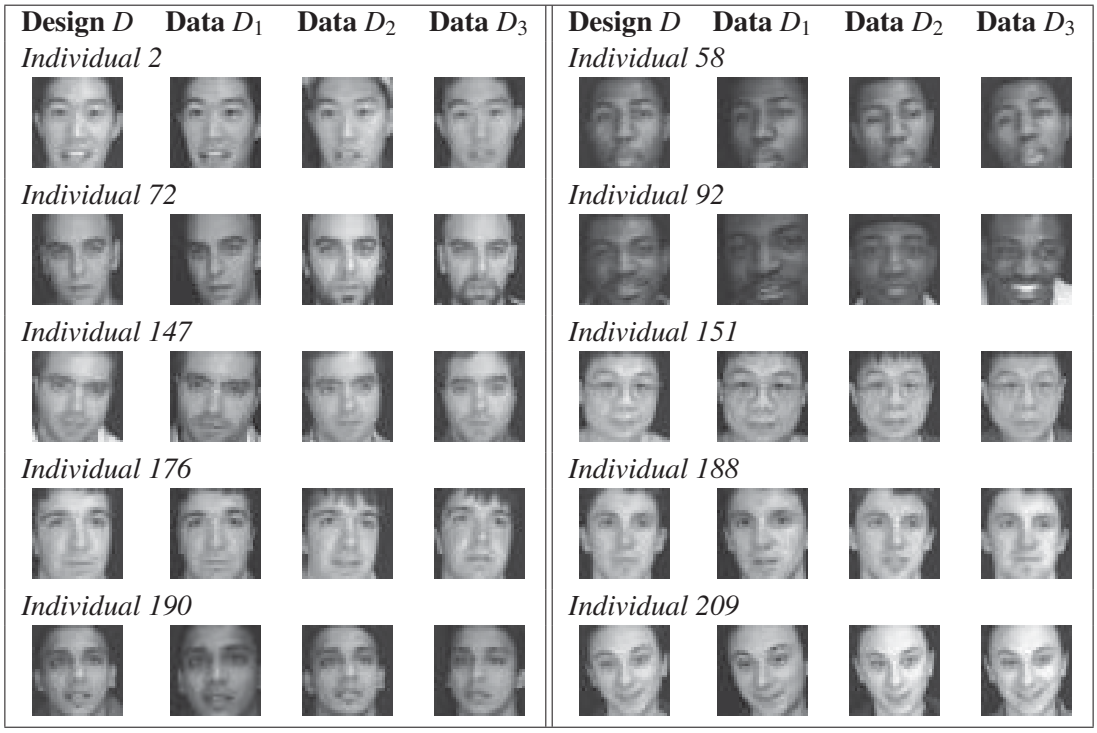


Figure 2.6 Sample images from individuals of interest detected in video sequences from the FIA database

Ten individuals of interest were selected, and one EoD was designed for each of them. Variants in expression, aging, pose, haircut, whiskers and beard made the problem more challenging (see Fig. 2.6). From the remaining individuals, 88 were selected to build the universal model (UM), and the rest were considered as unknown individuals and only appeared on the test datasets. Note that samples from individuals belonging to the UM do not appear in the test set, thus avoiding a positive bias.

One trajectory was retrieved from each individual in each capture session, and organized in four datasets. The total number of ROI samples contained in the trajectories from each design/update datasets is summarized on Table 2.2. As shown in the table, the CM is comprised of 9 trajectories from non-target individuals in the cohort, and the number of ROI samples is

different for each EoD. For instance, the CM of individual with FIA ID=2 is comprised of 1,746 reference ROI samples. ROI samples in the UM are 6,167, 2,966, and 3,188 are retrieved from each data block D_1, D_2 and D_3 respectively, divided in 88 non-target trajectories per block. Finally, the total number of ROI samples in the trajectories from unknown individuals is 10,240, 10,967, and 5,104 for D_1, D_2 and D_3 respectively. The fixed size memory containing the UM and CM is maintained with a first-in-first-out strategy, and it stores up to 12,000 facial regions belonging to the most recent trajectories from non-target individuals.

Table 2.2 Number of ROI samples in design and test trajectories for each individual of interest enrolled to the system

FIA Individual (k)	$\|T_k\|$, $T_k \in D$	$\|T_k\|$, $T_k \in D_1$	$\|T_k\|$, $T_k \in D_2$	$\|T_k\|$, $T_k \in D_3$
ID 2	149	114	109	119
ID 58	202	176	215	172
ID 72	223	144	184	151
ID 92	180	125	125	167
ID 147	235	128	163	161
ID 151	216	80	187	135
ID 176	113	90	210	126
ID 188	148	118	172	192
ID 190	190	132	92	88
ID 209	239	121	162	137

2.5.2 Implementation of the Proposed MCS

For proof-of-concept, the adaptive MCS proposed in Section 2.4 is implemented in the following way. The classification system in the MCS is formed of an adaptive EoD per individual (De-la Torre *et al.*, 2012a). The base classifier for the EoD is the Probabilistic Fuzzy ARTMAP (PFAM) (Lim and Harrison, 1995), which combines Fuzzy ARTMAP density estimation for learning category prototypes, with a non-parametric posterior probability distribution procedure inspired by the Probabilistic Neural Networks during the operational phase. A diversified pool of base classifiers is generated through a dynamic particle swarm optimization (DPSO) learning strategy (Connolly *et al.*, 2012). The DPSO learning algorithm was initialized with a swarm of 60 particles, 6 sub-swarms of maximum 5 particles, and a maximum of 30 iterations

(+5 to ensure convergence). The classifier corresponding to the global best particle, as well as the 6 local best classifiers from each sub-swarm are added to the ensemble. Finally, new classifiers are combined with previously trained ones ($\mathscr{P}_k$) using the Boolean combination (BC) that operates in the ROC space (Khreich *et al.*, 2010b). BC starts by regrouping classifiers according to performance and then combines all pairs of operations points for the two best classifiers, according to their representation in the ROC space. Then, the convex hull of the new operations points is successively combined with the next best classifiers, until the overall convex hull stops improving.

The CAMSHIFT is a well known kernel-based tracking algorithm that uses region-based features representation (Bradski, 1998). It uses a combination of a weighting kernel and a histogram to represent the target and attain frame-to-frame object tracks, using the probability distribution of faces in video. It dynamically handles the changing distributions by adjusting the size of the search window according to the area under such a window. The internal face representation consists of the skin probability histogram of the face, and the kernel is a simple step function. During data association, two histograms q_1 and q_2 corresponding to the predicted and actual facial regions respectively are compared with the Bhattacharyya coefficient given by:

$$Q_T \equiv \hat{B}(q_1, q_2) = \sum_{u=1}^{m} \sqrt{q_1(u) q_2(u, y)} \tag{2.6}$$

where u varies over all histogram bins, and y is the target position. Coefficient Q_t expresses the quality of a trajectory from one frame to another in terms of the similarity between predicted and actual face regions.

2.5.3 Experimental Protocol

Prior to computer simulations, four datasets were prepared using frontal videos of the FIA database. The design dataset D is comprised of the positive trajectories in the zoomed capture session 1. The adaptation datasets D_1 to D_3 are constructed with tracks from the un-

zoomed view of capture sessions 1 to 3, respectively. This capture scenario corresponds to an environment with gradual changes of face models due to aging. Negative samples are independently selected for each of the training/validation sets using Algorithm 2.4, by selecting samples from the CM and UM. Three different scenarios were prepared, with different design-update schemes.

- *Supervised learning on D only.* Considered a static system, designed on the first dataset D only. The test is performed on the other D_1 to D_3 datasets, but no update (additional learning) is performed. The performance in this scenario establishes the lower bound for the semi-supervised strategy, e.g., when no update is performed by the semi-supervised system. The approaches considered in this scenario include the TCM-kNN, a single PFAM, Learn++(PFAM) and EoD (PFAM).

- *Supervised incremental learning.* The system is first designed on D, and new reference samples become available (D_1 to D_3), and are incorporated after the test is performed. It is assumed that an expert has analyzed the video sequences of individuals enrolled to the system, and manually labels them in order to update the system. Adaptive approaches (PFAM$_{inc}$, Learn++(PFAM) and EoD$_{sup}$ (PFAM) $LTM_{KL,\lambda=\infty}$) were updated with only the new labeled data, and TCM-kNN is trained on batch mode, learning the past and new samples from scratch[5].

- *Partially-supervised learning.* Similarly to the supervised incremental learning scenario, the system is designed on D, and new information on test sessions D_1 to D_3 is incorporated when a trajectory T yields an accumulation curve that surpasses the update threshold, γ_k^u. The approaches considered in this scenario include the EoD$_{ss}$ (PFAM) with 6 different sizes of LTM: $\lambda = \{0, 25, 50, 75, 100, \infty\}$.

Learning is performed following 2x5-fold cross-validation for 10 independent experiments. Positive samples from the incoming trajectory are randomly and evenly split in 5 folds of the

[5]For a new block D_n, TCM-kNN must be trained from scratch using a data superset $D_{batch} = D \cup D_1 \cup ... \cup D_n$.

same size. The folds are first distributed in three different design sets, including two folds for training (D^t), $1\frac{1}{2}$ fold to stop training epochs (D^e), and $1\frac{1}{2}$ fold for fitness evaluation (D^f). Once the classifiers are trained, D^e and D^f are combined, randomized and divided into two equally distributed subsets to produce a validation data to estimate a fusion function (D^c), and to select the operations point (D^s). Negative samples are chosen from the UM as well as the CM according to CNN selection (Algorithm 2.4). In each training/validation dataset, 33% of positives is accompanied by approximately 58% of negatives from the UM, and the remaining 9% from the CM. About 87% of the negatives correspond to samples taken from the UM and 13% are from the cohort. This is expected, given that the superset D^- is composed of close to 13.63% of samples from the CM, and 86.37% of samples from the UM. The folds are distributed between the training/validation sets for each replication of the experiment, and average performance measures are produced with five different assignments. At replication 5, the sample order is randomized for each class and the five folds are regenerated. The procedure followed in each trial of the experiment is summarized in Algorithm 2.6.

Algorithm 2.6: Experimental protocol to evaluate each EoD_k, on a single 2×5 cross-validation trial

$D^- \Leftarrow UM \cup CM$ // Trajectories in the CM and UM
$EoD_k \Leftarrow DESIGN(T_k \in D, EoD_k \equiv \varnothing, LTM_k \equiv \varnothing, D^-)$ // Design the EoD_k with Algorithm 2.3
Estimate γ_k^d and γ_k^u using T_k and trajectories in D^-
for $t = 1...3$ **do**
  Evaluate performance of the EoD_k on D_t // Classifier and decision levels
  $D^- \Leftarrow UM \cup CM$ // Trajectories from CM and UM in D_t
  // For every trajectory in the new data block D_t
  for $T \in D_t$ **do**
    // If the accumulated predictions surpass the update threshold
    if $(A_k(T) \geq \gamma_k^u)$ **then**
      $T_k \Leftarrow T$ // Label the trajectory with tag k
      $EoD_k \Leftarrow UPDATE(T_k, EoD_k, LTM_k, D^-)$ // Update with T_k (Algorithm 2.3)
      Update γ_k^d and γ_k^d with T_k and trajectories in D^-

The proposed adaptive MCS was compared to other classifiers for FRiVS. The TCM-kNN was trained with a fixed $k = 1$ on a batch learning scheme, as followed in (Li and Wechsler, 2005). The Learn++ algorithm was initialized to generate 7 PFAM base classifiers on every incremental learning step, and weighted majority voting was validated on D^c. PFAM classifiers used in all other approaches were trained using a DPSO based learning strategy to optimize their hyperparameters.

2.5.4 Performance Analysis

The analysis of simulation results has been divided into three levels. First, *transaction-based analysis* shows the performance of the system based on classification decisions on each ROI. Then, a *subject-based analysis* allows a focus on specific individuals, which in turn allows for levels of performance depending on particular characteristics. Finally, a *trajectory-based analysis* shows the overall performance of the system (shown in Fig. 2.3), viewed by accumulating system predictions over input trajectories.

Transaction-based performance analysis is used to assess the performance of the system for matching ROI samples to facial models. The true positive rate (tpr) and false positive rate (fpr) are estimated for different (fpr,tpr) operational points, and connected to draw a receiver operations characteristic (ROC) curve. When equal priors and costs are assumed, the closest operations point to the upper-left corner corresponds to the optimal decision threshold. In applications with fpr constraint, the selection of the operations point is obtained from the graphical representation. The operations point is estimated on a validation subset used for operational predictions, providing a test (fpr,tpr) pair that reveals the generalization performance of the system at the selected point. The AUC (area under the curve) summarizes the performance depicted in a ROC graph, and the partial AUC (pAUC) focuses on a specific region of the curve, e.g. pAUC (5%) for an $fpr \leq 0.05$.

For different priors and costs of errors, the *Precision-Recall Operating Characteristic* (*PROC*) curve constitutes a graphical representation of detector performance where the impact of data

imbalance is considered. The precision between positive predictions ($precision = TP/(TP + FP)$) is combined with the tpr (or $recall$) to draw a PROC curve. In general, the tpr is increased when the amount of positive (minority class) samples augments. On the contrary, the *precision* decreases with this amount. To combine precision and recall at a particular operations point, the scalar F_1 produces a single performance indicator:

$$F_1 = 2 \cdot \frac{precision \cdot tpr}{precision + tpr} \tag{2.7}$$

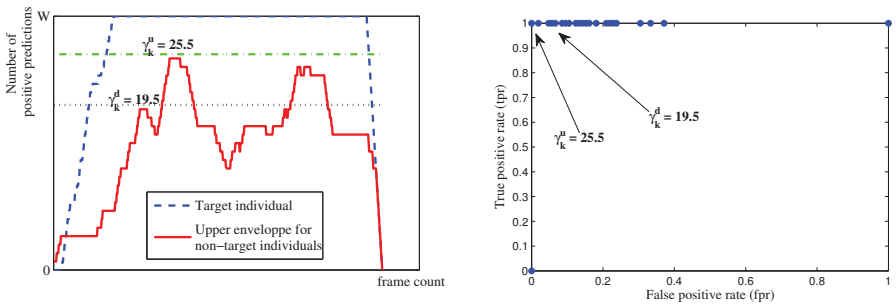
According to the "Doddington zoo" effect, the performance of biometric systems may vary drastically between individuals (Doddington *et al.*, 1998). Instead of using the overall amount of transactions, individual-specific error rates can be assessed according to four categories (types of animals). The resemblance of individuals performance to that of these animals can reveal fundamental weaknesses, and allows the development of more robust systems. According to this characterization, the system tend to perform well in a *sheep*-like individual, irrespective of whether this individual belongs to the target or non-target class. *Goat*-like individuals belong to the positive class, but are difficult to identify (low matching scores against themselves). A *wolf*-like individual belongs to the non-target class, and consistently impersonate different targets (high scores when matched against other individuals), and tend to elevate the false positive rate (fpr) of the system. Finally, a *lamb*-like individual belongs to the target class, and is easily impersonated (high matching scores when matched against others).

Table 2.3 Doddington's zoo thresholds for generalization performance at the operating point with $fpr = 1\%$, selected on validation data

Category	**Positive class**	**Negative class**
Sheep	$tpr \geq 50\%$ and not a lamb	$fpr \leq 1\%$
Lamb	At least 5% of non-target individuals are wolves	-
Goat	$tpr < 50\%$ and not a lamb	-
Wolf	-	$fpr > 1\%$

Typically, the likeliness of a user to one of the 4 aforementioned categories is defined at the score space. However, for binary classifiers, the confusion matrix can be used (Li and Wechsler, 2005). To establish a criterion, thresholds can be set at the fpr and fnr, and applied to each EoD_k. Table 2.3 shows a criterion based on a system constraint of $fpr \leq 1\%$, considering a good fnr when it is just below 50%.

Trajectory-based performance analysis allows to assess performance over time of the entire system for FRiVS (see Fig. 2.3). This analysis is specially relevant given that it provides a global performance assessment of the system for FRiVS, with combined impact of face segmentation, tracking, recognition and fusion. Thus, all system functions are employed to process a video stream, and decisions taken by an operator occur on a time scale longer than a frame rate. Within the decision fusion system, positive predictions of each EoD_k are accumulated over a moving window of time for input ROI samples that correspond to a high quality facial track. Assume for instance a system that produces predictions at a maximum of 30fps. Each detected ROI is presented to all user-specific EoDs of the system, which produces predictions (positive or negative) for each person enrolled to the system. Given a high quality face track, the number of positive predictions from an EoD should grow rapidly for the person of interest. Thus, the operator can more reliably detect a person of interest.



(a) Accumulation curves (b) ROC space

Figure 2.7 Trajectory-based analysis to evaluate the quality of a system for spatio-temporal FRiVS

The adaptive MCS proposed in this chapter accumulates the positive predictions (responses of each EoD_k) over a window of W predictions. As shown in Fig. 2.7a, the quality of this system can be evaluated graphically by observing the evolution of positive predictions according to the frame count (discrete time defined by the frame rate). In addition, once several individuals have appeared before of camera in a long video stream, and related trajectories have been processed, the quality of system decisions (i.e., the tpr, fpr, trr, frr) may be assessed over the range of decision threshold values, and represented in the ROC space (see Fig. 2.7b).

2.6 Results

2.6.1 Transaction-Based Analysis

Reference systems used in comparison reflect the current state-of-the-art approaches appearing in literature. TCM-kNN was proposed by Li and Wechsler in (Li and Wechsler, 2005), and constitutes a main reference in FRiVS. Learn++ is a popular reference point in ensemble-based techniques capable of supervised incremental learning (Polikar *et al.*, 2001). Modular architectures with a single classifier per individual have been used for FR in (Ekenel *et al.*, 2010), and implemented in experiments using monolithic PFAM and PFAM_{inc}. These modular architectures were extended to use ensembles of classifiers per individual in (Pagano *et al.*, 2012; Tax and Duin, 2008), and implemented in experiments as EoD (PFAM). In this research it is shown how the self update with the proposed approach presents higher level of performance with respect to those approaches that are not updated. And it may perform better than certain approaches that perform supervised incremental learning (e.g., Learn++), even though the proposed self update approach automatically assigns the labels to the trajectories in the update data.

Table 2.4 presents the average transaction-level performance for the 3 updating scenarios obtained after updating the proposed and reference systems on ROI samples from trajectories stored in data blocks D, D_1 and D_2 (while testing on D_1, D_2 and D_3, respectively). Systems are compared according to the partial AUC for a $0 \leq fpr \leq 0.05$: pAUC (5%), as well as fpr, tpr

and F_1 measures at a specific operating point selected on the validation ROC curve for a desired $fpr = 1\%$. Performance for modular systems were measured for each individual (user EoD), and average values are presented. In order to have comparable results for the multi-class TCM-kNN, empirical ROC curves were estimated on validation for each individual. The selection of the operations point, as well as performance evaluation were computed after applying the specialized rejection threshold of the TCM-kNN. Note that this rejection threshold is estimated on the training data, taking advantage of the peak-side-ratio that characterizes the distributions of p-values for each class.

Table 2.4 Average transaction-level performance of the system over the 10 individuals of interest and for 10 independent experiments. Systems were designed-updated with D, D_1 and D_2, and performance is shown after testing on D_1, D_2 and D_3 respectively (shown $D_1 \rightarrow D_2 \rightarrow D_3$). In all cases, the operations point was selected using the ROC space on the validation dataset D^s at a $fpr = 1\%$, except for the partial AUC that comprises the area for $0 \leq fpr \leq 0.05$. Bold values indicate significant differences from other approaches

fpr (%) $\downarrow$	**tpr (%)** $\uparrow$	$\mathbf{F_1}$ $\uparrow$	**pAUC (5%)** $\uparrow$
No update (supervised learning on D only)			
TCM-kNN			
20.13 ±0.42 → 24.74 ±0.50 → 18.88 ±0.53	**90.65 ±1.43** → **54.86 ±3.30** → 49.03 ±4.01	0.093 ±0.003 → 0.055 ±0.004 → 0.102 ±0.009	88.71 ±1.47 → 48.55 ±3.39 → 46.05 ±4.06
Monolithic PFAM			
0.95 ±0.18 → 0.94 ±0.20 → 0.82 ±0.18	80.84 ±2.05 → 32.88 ±3.44 → 37.35 ±3.91	0.665 ±0.019 → 0.280 ±0.029 → 0.358 ±0.035	90.40 ±1.21 → 54.67 ±3.24 → 61.54 ±3.58
Learn++ (PFAM)			
0.60 ±0.07 → 0.62 ±0.08 → 0.56 ±0.06	16.90 ±2.37 → 11.36 ±2.05 → 12.13 ±2.22	0.161 ±0.017 → 0.111 ±0.013 → 0.139 ±0.018	47.87 ±2.71 → 32.62 ±2.22 → 32.67 ±2.61
EoD (PFAM)			
0.62 ±0.09 → 0.64 ±0.10 → **0.53 ±0.09**	77.02 ±2.10 → 26.75 ±2.99 → 31.85 ±3.44	**0.679 ±0.018** → 0.255 ±0.025 → 0.337 ±0.032	**92.88 ±0.81** → 60.17 ±2.94 → 65.96 ±3.12
Supervised update (supervised incremental learning on $D \rightarrow D_1 \rightarrow D_2$)			
TCM-kNN			
20.13 ±0.42 → 22.81 ±0.41 → 18.32 ±0.19	**90.65 ±1.43** → 54.26 ±3.22 → **87.91 ±1.67**	0.094 ±0.003 → 0.058 ±0.004 → 0.175 ±0.004	88.71 ±1.47 → 48.54 ±3.34 → 83.16 ±2.29
PFAM_{inc}			
0.95 ±0.18 → 1.20 ±0.12 → 1.91 ±0.24	80.84 ±2.05 → 54.06 ±3.46 → 84.52 ±2.31	0.665 ±0.019 → **0.438 ±0.029** → 0.666 ±0.024	90.40 ±1.21 → 69.18 ±2.86 → 87.75 ±1.66
Learn++ (PFAM)			
0.60 ±0.07 → **0.57 ±0.04** → 1.19 ±0.11	16.90 ±2.37 → 11.87 ±1.80 → 20.57 ±2.78	0.161 ±0.017 → 0.128 ±0.014 → 0.192 ±0.020	47.87 ±2.71 → 36.81 ±2.45 → 34.19 ±2.64
EoD_{sup} (PFAM) $\text{LTM}_{KL,\lambda=\infty}$			
0.62 ±0.09 → 0.67 ±0.05 → 0.84 ±0.07	77.02 ±2.10 → 45.51 ±3.63 → 76.70 ±2.71	**0.679 ±0.018** → 0.404 ±0.031 → **0.691 ±0.023**	**92.88 ±0.81** → **72.03 ±2.76** → **93.64 ±0.84**
Self update (semi-supervised incremental learning on $D \rightarrow D_1 \rightarrow D_2$)			
EoD_{ss} (PFAM) $\text{LTM}_{KL,\lambda=\infty}$			
0.62 ±0.09 → 0.74 ±0.07 → 0.93 ±0.11	77.02 ±2.10 → 43.33 ±3.59 → 50.10 ±4.12	0.679 ±1.77 → 0.388 ±0.031 → 0.461 ±0.037	92.88 ±0.81 → 68.50 ±2.90 → 75.60 ±3.04

In the no-update scenario, the EoD (PFAM) approach is generally the most accurate approach in terms of pAUC (5%). Overall results for all approaches show a degradation in the system performance after testing on D_2, with a slight recovery after testing on D_3, indicating the presence of changes in the classification environment going from D to D_1 and to D_2. This decline in performance underscores the importance of adapting facial models as new reference videos become available.

At the selected operations point (fpr=1%), it is interesting to note that, compared to monolithic classifiers (PFAM and TCM-kNN), both ensemble-based classifiers provide lower fpr, along with a lower standard error. The only multi-class classifier used in the comparison, the TCM-kNN, yields a significantly higher fpr, even though it was designed to avoid false acceptances by using a specialized rejection threshold. This issue is related to the difficulty faced by multi-class classifiers in estimating multiple decision boundaries during the same design process: between cohort and unknown individuals, and between individuals in the cohort. Modular architectures simplify the task by optimizing parameters for user-specific 2-class classifiers for determining individual-specific bounds, which provides greater discrimination when design data per target individual is limited (Oh and Suen, 2002). Consequently, TCM-kNN achieves the highest tpr, but fails meeting constraints for the fpr on test data. Ensemble approaches (Learn++ and EoD) have the lower fpr, although the PFAM and EoD (PFAM) provide the highest tpr and F_1 measures. This translates to a greater discrimination for target ROI samples. Results suggest that the EoD (PFAM) can achieve the most robust overall performance to gradually changing environments.

The average results (Table 2.4) for the supervised update scenario show the impact on performance of updating the facial models. The degradation seen in the no-update case is reduced. The pAUC (5%) reveals that the EoD$_{sup}$ (PFAM) LTM$_{KL,\lambda_k=\infty}$ provides a significantly higher level of performance, which confirms the utility of adaptive ensembles. This approach establishes an upper bound for self-updating, given that it correctly updates facial models with every new target trajectory. As in the no-update case, it can be seen that adaptive ensembles present lower fpr but also lower tpr, and PFAM$_{inc}$ and EoD$_{sup}$ (PFAM) LTM$_{KL,\lambda=\infty}$ provide

the greater discrimination on target ROI samples. TCM-kNN presents the most significant degradation in performance after testing on D_2, even though it was retrained with samples from $D \cup D_1$. However, it also presents an important recovery after testing on D_3. A Kruskall-Wallis statistical test on the pAUC (5%) between the EoD$_{sup}$ (PFAM) and PFAM$_{inc}$ gives a p-value of 0.0123, which confirms that the differences between the mean performances are significant with a 95% confidence interval.

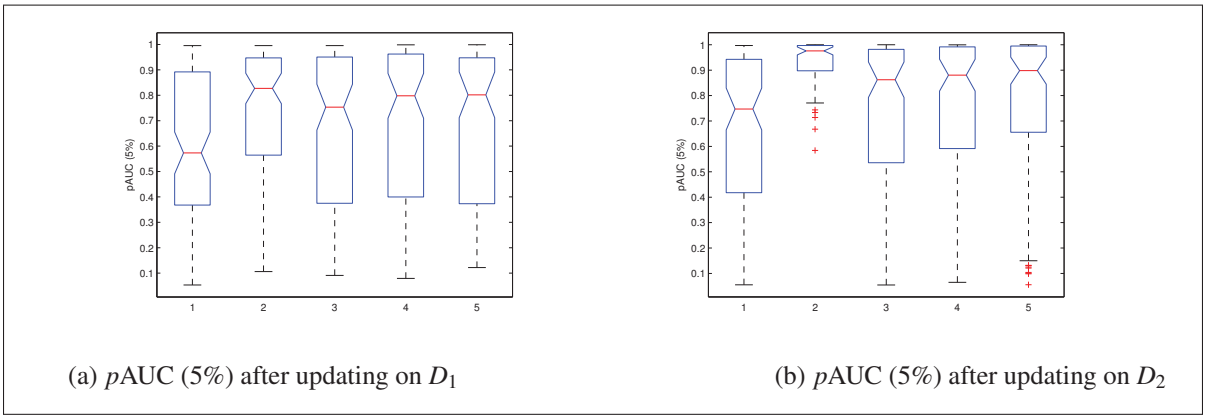
Average results achieved with the proposed semi-supervised adaptive MCS (EoD$_{ss}$) indicate that the performance is generally comparable to that of the supervised approaches in terms of pAUC (5%), although a higher fpr is eventually present. This degradation is the cumulative effect of false adaptations followed by trajectories that are incorrectly labeled (see analysis in Section 2.6.2). However the performance of the semi-supervised system evolves with a general improvement with respect to the no-update case as new reference data is integrated. And it remains close to the upper bound established by the approaches that perform supervised update.

Table 2.5 Average transaction-level performance of the EoD$_{ss}$ (PFAM) system given different LTM sizes λ_k, after testing on $D_1 \rightarrow D_2 \rightarrow D_3$. In all cases, the operations point was selected using the ROC space on the validation dataset D^s for an $fpr = 1\%$, except for the pAUC (5%) that comprises the area for $0 \leq fpr \leq 0.05$

fpr % ↓	**tpr % ↑**	**F$_1$ ↑**	**pAUC (5%) ↑**
EoD$_{ss}$ (PFAM), LTM$_{KL,\lambda=0}$			
0.62 ±0.09 → 0.96 ±0.09 → 1.55 ±0.22	77.02 ±2.10 → 44.25 ±3.60 → 51.39 ±3.95	0.679 ±0.018 → 0.373 ±0.030 → 0.428 ±0.032	92.88 ±0.81 → 65.88 ±2.92 → 72.37 ±3.09
EoD$_{ss}$ (PFAM), LTM$_{KL,\lambda=25}$			
0.62 ±0.09 → 1.42 ±0.22 → 1.74 ±0.24	77.02 ±2.10 → 36.17 ±3.43 → 48.83 ±3.85	0.679 ±0.018 → 0.306 ±0.029 → 0.402 ±0.031	92.88 ±0.81 → 62.80 ±3.04 → 70.95 ±3.05
EoD$_{ss}$ (PFAM), LTM$_{KL,\lambda=50}$			
0.62 ±0.09 → 1.25 ±0.16 → 1.44 ±0.15	77.02 ±2.10 → 35.28 ±3.35 → 48.84 ±3.90	0.679 ±0.018 → 0.304 ±0.029 → 0.407 ±0.032	92.88 ±0.81 → 62.35 ±3.08 → 71.48 ±3.13
EoD$_{ss}$ (PFAM), LTM$_{KL,\lambda=75}$			
0.62 ±0.09 → 1.27 ±0.16 → 1.90 ±0.29	77.02 ±2.10 → 36.76 ±3.53 → 50.13 ±3.90	0.679 ±0.018 → 0.307 ±0.029 → 0.404 ±0.032	92.88 ±0.81 → 61.50 ±3.12 → 71.84 ±3.11
EoD$_{ss}$ (PFAM), LTM$_{KL,\lambda=100}$			
0.62 ±0.09 → **0.92 ±0.09** → **1.45 ±0.18**	77.02 ±2.10 → **45.43 ±3.71** → **54.27 ±3.86**	0.679 ±0.018 → **0.385 ±0.031** → **0.468 ±0.033**	92.88 ±0.81 → **68.44 ±3.00** → **74.93 ±2.98**

A key parameter related to the accuracy and resources of EoD$_{ss}$ (PFAM) systems is the LTM size needed to store validation data. Table 2.5 shows the evolution of the average performance

for LTM sizes $\lambda_k = \{0, 25, 50, 75, 100\}$ patterns. As the system self-updates, the overall performance improves when λ_k grows, at the expense of memory and computational complexity. However, this trend occurs differently for distinct individuals, as analyzed in the subject-based analysis. Finally, Fig. 2.8 shows the box plots for pAUC (5%) for the EoD$_{ss}$ (PFAM) system with different λ_k values. The first box in the graphs corresponds to the EoD (PFAM) that learns only on D, and establishes the lower bound in performance. The second box is the supervised EoD$_{sup}$ (PFAM) with a $\lambda_k = \infty$, and establishes the upper bound. It can be seen that pAUC (5%) grows with the LTM size. Using a $\lambda_k = 100$ provides a performance that is comparable to what is seen when $\lambda_k = \infty$.



(a) pAUC (5%) after updating on D_1

(b) pAUC (5%) after updating on D_2

Figure 2.8 Box plots comparing the pAUC (5%) of systems (a) after learning D_1 (testing on D_2), and (b) after learning D_2 (testing on D_3). The systems from left to right are (1) EoD (PFAM), (2) EoD$_{sup}$ (PFAM) LTM$_{KL,\lambda_k=\infty}$, (3) EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda_k=0}$, (4) EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda_k=100}$, (5) EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda_k=\infty}$

2.6.2 Subject-Based Analysis

Table 2.6 presents the average performance of ensembles for the semi-supervised scenario obtained after self-update using ROI samples from trajectories stored in D, D_1 and D_2. The LTM size used corresponds to $\lambda_k = 25$ and 100 patterns. Modules 58 and 209 correspond to individuals of interest with good initial performance (pAUC (5%) ≥ 95). They are easy to detect with an EoD$_{ss}$ (PFAM) ($tpr \geq 50\%$), and to differentiate from non-target individuals ($fpr \leq 1\%$):

These are typically *sheep*-like individuals in the Doddington zoo taxonomy. Results after learning D reveal the existence of 4 non-target individuals that are incorrectly detected more than 1% of the time (*wolves*) in both cases, corresponding to the 2.58% of the non-target individuals during tests. In contrast, EoDs$_{ss}$ 151 and 188 were selected because they initially provide poor performance (pAUC (5%) $< 95\%$). EoD$_{ss}$ 151 corresponds to an individual that is difficult to detect by the system ($tpr < 50\%$), but is also difficult to impersonate ($fpr \leq 1\%$). The test at $t = 1$ reveals 5 wolves for this *goat*-like individual in the Doddington zoo taxonomy. The number of wolves corresponds to 3.23% of non-target individuals. EoD$_{188}$ corresponds to an individual which while being easy to detect by the system ($tpr \geq 50\%$), it is also easy to impersonate ($fpr > 1\%$). The test on D_1 reveals 32 wolves, corresponding to 20.65% of non-target individuals. Given the number of wolves, EoD$_{188}$ corresponds to a *lamb*-like individual.

Results for EoD$_{ss}$ 58 after updating on D_1 (testing on D_2) show a decline in pAUC (5%) performance for both λ_k values. However, the F_1 performance shows a greater decline for $\lambda_k = 100$, which reveals that D_1 contains some ROI samples that corrupt the facial model, and degrades the EoD$_{ss}$ (PFAM) accuracy. It can be seen however that some of these are filtered out by the KL selection strategy, given the higher performance with $\lambda_k = 25$. The overall results suggests that for this sheep-like individual, the performance can be maintained using small λ_k values.

The pAUC (5%) for EoD$_{ss}$ 209 after testing on D_2 also shows a decline in performance for $\lambda_k = 25$. Alto a small recovery is shown after testing on D_3, performance does not regain the same level due to the lack of representative validation data. On the other hand, an LTM with $\lambda_k = 100$ is shown to be able to maintain and improve the level of performance. This results suggest that sheep-like individuals benefit from higher λ_k values, and low λ_k values may lead to the corruption of the facial models. Given the results form EoDs 58 and 209, one can conclude that high values of λ_k ensure performance for sheep-like individuals, and individual-specific λ_k values should be estimated based on the evolution of specific EoDs.

Table 2.6 Average performance of the system for 4 individuals of interest over 10 independent experiments, after test on $D_1 \rightarrow D_2 \rightarrow D_3$. Two cases that initially provide a high level of performance correspond to EoDs with an initial pAUC (5%) $\geq$ 95% on D_1. Cases with initial performance that is poor are those with an initial pAUC (5%) $<$ 95% on D_1

EoDs with good initial performance		
	EoD$_{ss}$ 58 (*sheep*-like)	EoD$_{ss}$ 209 (*sheep*-like)
EoD$_{ss}$ (PFAM), semi-supervised incremental learning, LTM_{KL}, $\lambda = 25$		
fpr (%) $\downarrow$	0.23 ±0.09 $\rightarrow$ 0.85 ±0.07 $\rightarrow$ 1.46 ±0.45	0.34 ±0.07 $\rightarrow$ 5.44 ±1.56 $\rightarrow$ 2.74 ±0.68
tpr (%) $\uparrow$	84.43 ±3.33 $\rightarrow$ 39.35 ±7.06 $\rightarrow$ 44.24 ±12.73	86.28 ±3.54 $\rightarrow$ 11.79 ±9.76 $\rightarrow$ 33.80 ±13.22
F$_1$ $\uparrow$	0.849 ±0.023 $\rightarrow$ 0.402 ±0.061 $\rightarrow$ 0.373 ±0.077	0.792 ±0.018 $\rightarrow$ 0.047 ±0.031 $\rightarrow$ 0.205 ±0.086
pAUC (5%) $\uparrow$	98.45 ±0.23 $\rightarrow$ 73.74 ±3.52 $\rightarrow$ 79.52 ±5.93	97.61 ±0.31 $\rightarrow$ 46.81 ±10.51 $\rightarrow$ 64.13 ±10.52
EoD$_{ss}$ (PFAM), semi-supervised incremental learning, LTM_{KL}, $\lambda = 100$		
fpr (%) $\downarrow$	0.23 ±0.09 $\rightarrow$ 0.86 ±0.09 $\rightarrow$ 1.62 ±0.39	0.34 ±0.07 $\rightarrow$ 0.46 ±0.07 $\rightarrow$ 1.10 ±0.26
tpr (%) $\uparrow$	84.43 ±3.33 $\rightarrow$ 35.44 ±8.10 $\rightarrow$ 51.16 ±14.32	86.28 ±3.54 $\rightarrow$ 88.33 ±3.33 $\rightarrow$ 98.10 ±0.71
F$_1$ $\uparrow$	0.849 ±0.023 $\rightarrow$ 0.353 ±0.066 $\rightarrow$ 0.384 ±0.093	0.792 ±0.018 $\rightarrow$ 0.793 ±0.023 $\rightarrow$ 0.802 ±0.037
pAUC (5%) $\uparrow$	98.45 ±0.23 $\rightarrow$ 74.58 ±3.54 $\rightarrow$ 80.44 ±6.34	97.61 ±0.31 $\rightarrow$ 97.16 ±0.27 $\rightarrow$ 99.59 ±0.11
EoDs with bad initial performance		
	EoD$_{ss}$ 151 (*goat*-like)	EoD$_{ss}$ 188 (*lamb*-like)
EoD$_{ss}$ (PFAM), semi-supervised incremental learning, LTM_{KL}, $\lambda = 25$		
fpr (%) $\downarrow$	0.13 ±0.04 $\rightarrow$ 0.43 ±0.22 $\rightarrow$ 0.31 ±0.15	2.54 ±0.57 $\rightarrow$ 0.95 ±0.09 $\rightarrow$ 0.57 ±0.21
tpr (%) $\uparrow$	37.50 ±7.91 $\rightarrow$ 19.14 ±10.17 $\rightarrow$ 51.19 ±13.86	89.58 ±4.26 $\rightarrow$ 85.17 ±4.68 $\rightarrow$ 90.78 ±5.33
F$_1$ $\uparrow$	0.447 ±0.065 $\rightarrow$ 0.182 ±0.089 $\rightarrow$ 0.509 ±0.112	0.472 ±0.054 $\rightarrow$ 0.670 ±0.024 $\rightarrow$ 0.863 ±0.039
pAUC (5%) $\uparrow$	82.19 ±5.46 $\rightarrow$ 65.30 ±9.46 $\rightarrow$ 91.34 ±3.85	91.12 ±2.41 $\rightarrow$ 95.48 ±1.14 $\rightarrow$ 99.73 ±0.05
EoD$_{ss}$ (PFAM), semi-supervised incremental learning, LTM_{KL}, $\lambda = 100$		
fpr (%) $\downarrow$	0.13 ±0.04 $\rightarrow$ 0.25 ±0.14 $\rightarrow$ 0.26 ±0.15	2.54 ±0.57 $\rightarrow$ 1.18 ±0.20 $\rightarrow$ 0.31 ±0.10
tpr (%) $\uparrow$	37.50 ±7.91 $\rightarrow$ 27.17 ±12.63 $\rightarrow$ 48.15 ±13.56	89.58 ±4.26 $\rightarrow$ 89.88 ±3.09 $\rightarrow$ 93.70 ±1.74
F$_1$ $\uparrow$	0.447 ±0.065 $\rightarrow$ 0.274 ±0.119 $\rightarrow$ 0.498 ±0.112	0.472 ±0.054 $\rightarrow$ 0.667 ±0.032 $\rightarrow$ 0.920 ±0.013
pAUC (5%) $\uparrow$	82.19 ±5.46 $\rightarrow$ 68.64 ±9.32 $\rightarrow$ 91.39 ±3.86	91.12 ±2.41 $\rightarrow$ 96.39 ±0.48 $\rightarrow$ 99.72 ±0.05

With EoD$_{ss}$ 151, pAUC (5%) and F_1 performance declines after testing on D_2. This decline accentuated when $\lambda_k = 25$ patterns. Similarly to EoD$_{ss}$ 58, this trend reveals that D_2 contains some samples that corrupt this facial model. However, in this case, the system benefits from

higher λ_k values. Both EoDs show an increase in performance after testing on D_3, showing comparable performance in terms of F_1 and pAUC (5%) for both λ_k values. This reveals that, in the presence of corrupted data, goat-like individuals benefit from greater LTM sizes.

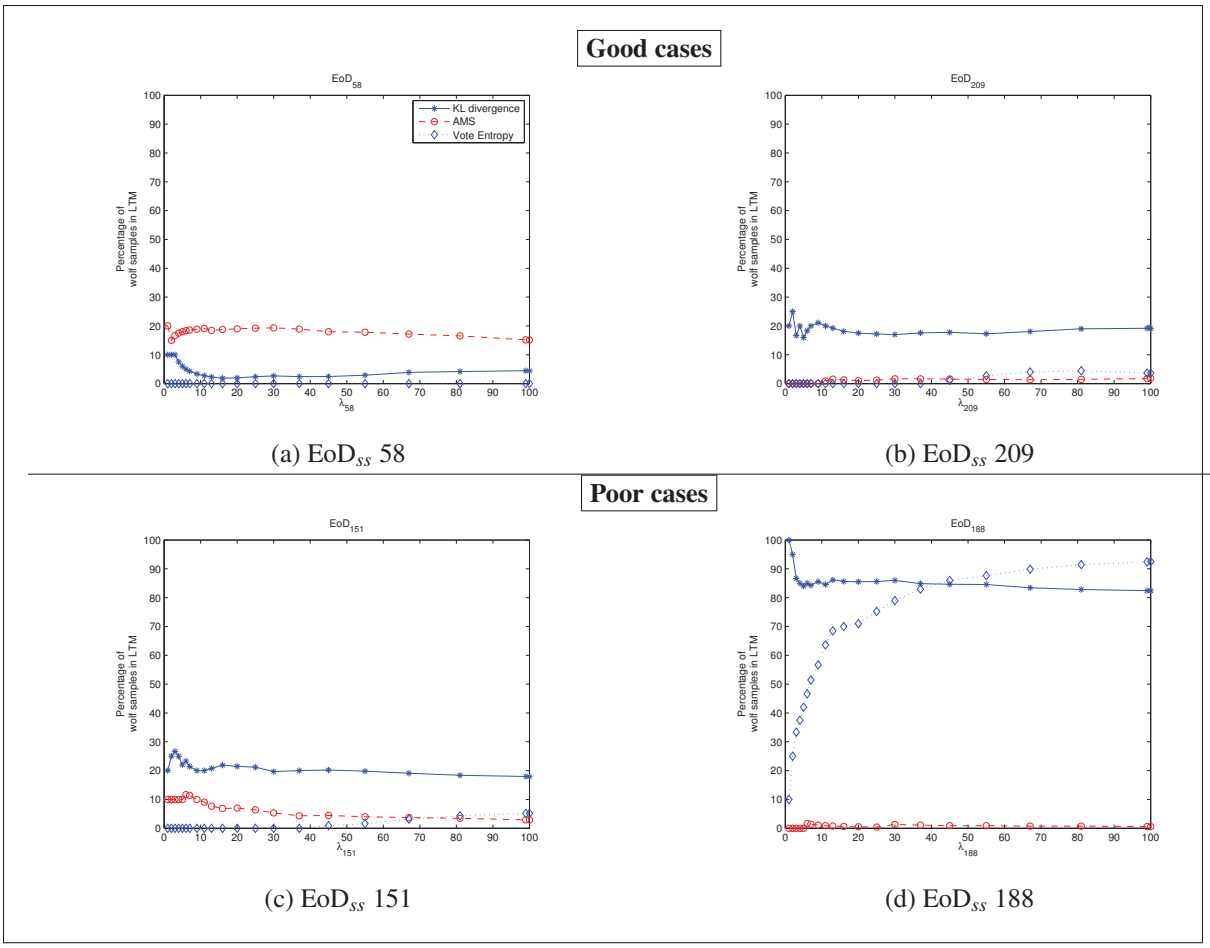
EoD$_{ss}$ 188 presents a constant increase in pAUC (5%) and F_1 performance. Despite the number of incorrect updates produced by multiple wolves, the fpr decreases after each self-update. This suggests that lamb-like individuals benefit from diverse samples from these updates as well. Similar performance is achieved by the EoD$_{ss}$ (PFAM) for small or large λ_k values.

It is well known that samples from wolf-like individuals negatively affect the fpr of EoDs, and by definition, the effect is more pronounced if the EoD corresponds to a lamb-like individual. Figure 2.9 presents the percentage of samples from wolf-like individuals selected by KL divergence, Average Margin Sampling (AMS) and Vote Entropy (VE), corresponding to the analyzed individuals of interest. Different sizes of LTM were tested following the exponential scale $\lambda_k = \lceil e^x \rceil$, where $x = 0, 0.2, 0.4, ...4.6$ [6]. Results show no clear tendency for the good cases, as shown in the graphs in Figure 2.9**a** and 2.9**b**. For these two sheep-like individuals (EoD_{58} and EoD_{209}) the AMS and KL divergence select a similar amount of samples from wolf-like individuals in different cases. As shown in Figure 2.9**c**, the KL divergence retrieves more samples from wolf-like individuals when the EoD corresponds to a goat-like individual. Finally, Figure 2.9**d** shows that for lamb-like individuals, the KL divergence is specially effective in finding samples from wolf-like individuals given a small LTMs ($\lambda < 50$). In summary, the KL divergence is useful in cases with poor initial performance (lamb-like and goat-like individuals), and with only small LTM sizes.

2.6.3 Trajectory-Based Analysis

Fig. 2.10 presents the accumulation curves showing the positive predictions produced by the EoDs in response to target and non-target trajectories in D_1 (replication 1). The detection and update thresholds estimated on the validation set are also depicted on the graphs. As can be

[6] Note that $\lambda_k = \lceil e^{4.6} \rceil = 100$, the maximum λ_k considered in experiments.



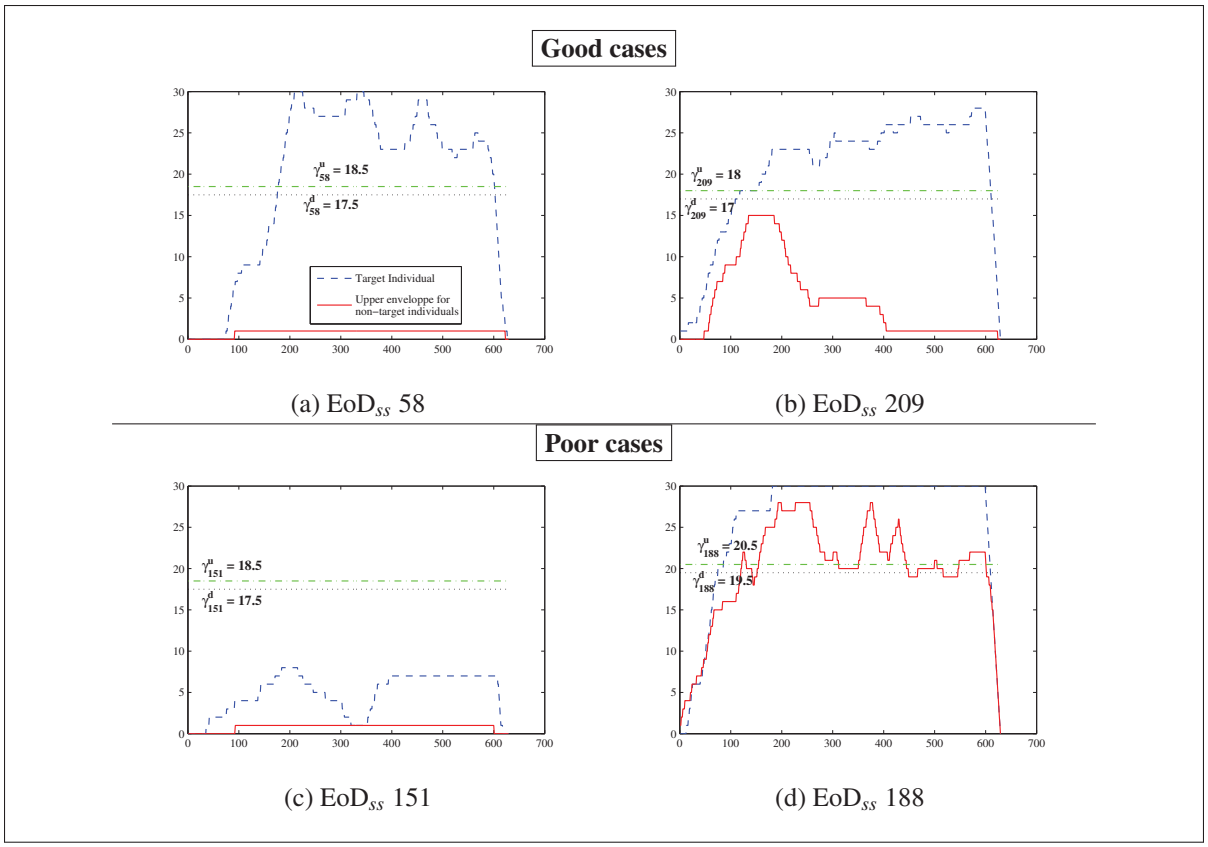
(a) EoD$_{ss}$ 58 (b) EoD$_{ss}$ 209

(c) EoD$_{ss}$ 151 (d) EoD$_{ss}$ 188

Figure 2.9 Percentage of wolf-like individuals in LTMs for the EoDs in the subject-based analysis

observed in this case, the accumulative curves corresponding to the two sheep-like individuals surpass both detection and update thresholds. And the upper envelope for non-target individuals is always below the thresholds, which means that none of the negative trajectories was incorrectly assigned to the target individual. EoDs for IDs 58 and 209 both exhibit a correct detection through D_1, allowing for the correct rejection of all negative trajectories in D_1.

The accumulative curves for EoD$_{ss}$ 151 and 188 for the same replication are also presented in Fig. 2.10. While the goat-like individual (ID 151) remains hard to detect, the lamb-like individual (ID 188) is impersonated by wolves present in D_1. Results suggest that the level of Γ_k (in Eq. 2.3) should be different for each type of individual. For instance, sheep-like



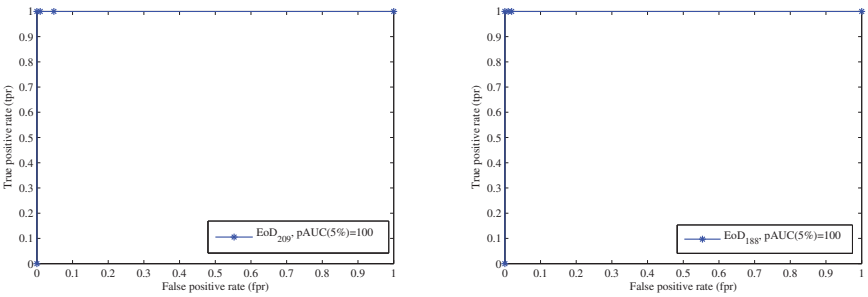
(a) EoD$_{ss}$ 58 (b) EoD$_{ss}$ 209

(c) EoD$_{ss}$ 151 (d) EoD$_{ss}$ 188

Figure 2.10 Accumulated positive prediction curves produced by the EoD$_{ss}$ (PFAM) of target vs. the non-target individuals, after training on D (testing on D_1), along with detection and update thresholds

individuals require smaller Γ_k values, and lamb-like individuals require larger Γ_k values. On the other hand, goat-like individuals may require a reduction of the detection threshold.

Fig. 2.11 shows the ROC curves for the overall system at the decision level. These curves were obtained by varying the decision thresholds on the accumulation curves produced by target and non-target trajectories in D_3 (Fig. 2.10). It shows the high level of discrimination achieved with these EoD$_{ss}$ (PFAM) at the decision fusion system after two updates, by accumulating evidence. Even though the selected update threshold γ^{μ}_{188} permitted some false updates after testing on D_1, the EoD$_{ss}$ increased its level of discrimination, achieving only correct updates after testing on D_3.



(a) ROC curve for EoD$_{ss}$ 209 (b) ROC curve for EoD$_{ss}$ 188

Figure 2.11 ROC curves for EoDs 209 (a) and 188 (b) at the decision fusion level, test on D_3, experiment trial 1. In both cases the final curves are perfect after two updates, even though the EoD$_s s$ 188 was updated 5 times with non-target trajectories in D_1

Table 2.7 The average performance of the overall system following a trajectory-based analysis. The number of target trajectories is 10, and the number of non-target trajectories is 1050 for the 10 replications after test on D_1. Results are produced by the system EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda_k=100}$, for the 4 cases in analysis

Measure	**EoDs with good initial performance**		**EoDs with bad initial performance**	
	EoD$_{ss}$ 58	EoD$_{ss}$ 209	EoD$_{ss}$ 151	EoD$_{ss}$ 188
tpr	100.00	100.00	50.00	100.00
fpr	0.00	0.00	0.00	0.86
F$_1$	1.00	1.00	0.667	0.6896
pAUC (5%)	100.00	100.00	51.25	91.40

Table 2.7 shows the average number of correct and incorrect trajectories detected by the selected EoD$_{ss}$ (PFAM) at the decision level. The benefit of accumulating predictions over a trajectory becomes evident for these EoDs by comparing the tpr and fpr before and after decision fusion. For instance, EoD$_{ss}$ 58 presents a $tpr = 84.43\%$ and $fpr = 0.23\%$ using transaction-based decisions (see Table 2.6), but using the whole trajectories in making the decision it produces a $tpr = 100\%$ and $fpr = 0\%$. This means that every time a target trajectory from D_1 was presented to the system, it was correctly detected by the corresponding EoD$_{ss}$, and all non-target trajectories were correctly rejected. A similar behavior is shown by EoD$_{ss}$ 209, which confirms that EoDs for sheep-like individuals may achieve a high level of discrimination with the proposed approach.

Performance is also seen to be increasing in EoDs for individuals 151 and 188, the tpr growing considerably and simultaneously reducing the fpr to about 0%. Moreover, using the decisions based on trajectories, the number of wolves is reduced from 32 to only 5 for the $wolf$-like individual 188. This suggests that the EoDs for both goat- and lamb-like individuals may also benefit from the proposed trajectory-based decision scheme.

Table 2.8 IDs corresponding to the trajectories in FIA that surpassed the update threshold and were used for updating the selected EoDs on different replications (r) of the experiment (EoD_{ss}, $\text{LTM}_{KL,\lambda_k=100}$). Bold numbers correspond to trajectories used for correct updates, and conflicts are marked with a box around the ID of the trajectory

Rep.	EoD_{ss} 58	EoD_{ss} 151	EoD_{ss} 188	EoD_{ss} 209	EoD_{ss} 58	EoD_{ss} 151	EoD_{ss} 188	EoD_{ss} 209
	Update trajectories in D_1				Update trajectories in D_2			
r=1	**58**	-	6,60,186,**188**,193,224	**209**	-	-	**188**	**209**
r=2	**58**	-	**188**,224	**209**	-	-	104,**188**	-
r=3	**58**	**151**	**188**	**209**	**58**	-	-	**209**
r=4	**58**	-	**188**	**209**	-	-	-	**209**
r=5	**58**	-	**188**,224	**209**	**58**,134	-	**188**	**209**
r=6	**58**	**151**	**188**	**209**	**58**	**151**	104,**188**	**209**
r=7	**58**	-	**188**,224	**209**	-	-	**188**	**209**
r=8	**58**	**151**	**188**	**209**	**58**	-	104,122,**188**	**209**
r=9	**58**	**151**	**188**,224	**209**	-	**151**	104,151,153,**188**	**209**
r=10	**58**	**151**	**188**	**209**	**58**	**151**,174	104,**188**	**209**

Table 2.8 provides further details on the updates over replications 1 to 10 for selected EoDs with $\text{LTM}_{KL,\lambda_k=100}$. After testing on D_1, EoD_{ss} 58 is always correctly and never incorrectly updated. However, after testing on D_2, only 50% of correct updates were performed, and an incorrect update was present at replication 5. This phenomenon is explained by the drop in performance due to the existence of ROI samples on D_1 that corrupted the facial model, as discussed earlier. A similar trend is presented by EoD_{ss} 151, dropping from 5 correct updates on D_1, to 3 correct and 1 incorrect updates. However, at replication 9, the correct update is discarded due to the conflict with EoD_{ss} 188. The facial model for individual 188 was correctly updated on all replications after testing on D_1, but 9 wrong updates were also performed on five of the replications. After test on D_2 the number of correct updates dropped to 8, and incorrect updates dropped to 8, in 5 of the replications. And one of the incorrect updates was discarded due to the conflict detected with EoD_{ss} 151. A different trend is shown by EoD_{ss} 209, for

which a reduction in the number of correct updates was only seen at replication 2, and never presented a wrong update.

2.7 Conclusion

In this chapter, an adaptive MCS is proposed for video-to-video FR, where the face of each target individual is modeled using an ensemble of 2-class classifiers. During operations, this new system integrates information from a face tracker and individual-specific ensembles for robust spatio-temporal recognition and for efficient self-update of facial models. The tracker defines a facial trajectory for each individual that appears in a video. Spatio-temporal FR occurs if the number of positive predictions accumulated along a trajectory surpass the detection threshold for an individual-specific ensemble. A higher update threshold allows the system to determine if the trajectory incorporates enough confidence for self-update of facial models. To update a facial model, all target samples extracted from the trajectory are combined with non-target samples selected from the cohort and universal models. Facial models are updated using a learn-and-combine strategy to avoid knowledge corruption that can occur during self-update with an incremental learning classifier. In addition, a memory management strategy based on Kullback-Leibler divergence is used to rank and select the most relevant target and non-target reference ROI samples for validation.

Proof of concept validation has been performed on the CMU-FIA video dataset with a particular realisation of the proposed system.The individual-specific EoDs are formed with of ARTMAP neural network classifiers generated using a DPSO incremental learning strategy, where classifiers are combined using BC. Transaction-level results indicate that the proposed adaptive MCS improved pAUC (5%) by about 8% over the system that do not perform self-update. It provides an average performance comparable to the same system that performs supervised update of facial models with all relevant trajectories. Subject-level analysis reveals that facial models from sheep- and goat-like individuals benefit from using a large LTM, while *lamb*-like individuals present similar performance with large or small LTM sizes. This is a consequence of the capacity of the KL divergence to select samples from wolf-like individuals,

which are more numerous for EoDs corresponding to *lamb*-like individuals. For trajectory-level analysis shown by the accumulated decisions, the system increases discrimination and robustness compared to transaction-level decisions. In all the cases that were analyzed, the individual-specific EoDs were able to simultaneously increase the overall pAUC (5%), tpr and F_1 measures, and reduce the fpr. Finally, an analysis of the updates achieved by the system shows that by virtue of the increased discrimination, it presented a low number of incorrect updates even with the large number of non-target trajectories presented to the system during simulations.

In this chapter, trajectories define the design samples used for (re)enrollment (supervised learning) and update (supervised or unsupervised learning) of facial models encoded in a video-to-video FR system. The proposed MCS has been characterized using data that exhibits a gradual pattern of changes over different capture sessions. Future research should analyze performance under abrupt patterns of change, as seen in sharp variations of illumination and face pose. A dynamic adaptation of the fusion functions of the ensembles to these scenarios may allow a better exploitation of the availability of abundant operational data. Since the proportion of target to non-target ROIs captured in practice is imbalanced, and the level of imbalance changes over time, classifier ensembles should be selected dynamically according to the context to improve performance. Regarding resource management, the exploration of pruning strategies for ensembles is another open issue. In practice, the system should exploit internal knowledge (age, performance relevance, etc.) to remove some older or redundant classifiers over time. With respect to the KL based LTM management scheme, it might be characterized on different applications of adaptive ensembles, like iris or gait recognition, signature verification, or in general object recognition. Finally, the system may also benefit from knowledge of ROI samples from wolf- and goat-like individuals, and the amount of validation samples stored in LTM may be optimized per individual. This could allow to select target and non-target ROI samples that lead to more discriminant individual-specific EoDs.

CHAPTER 3

AN ADAPTIVE ENSEMBLE-BASED SYSTEM FOR FACE RECOGNITION IN PERSON RE-IDENTIFICATION

Miguel De-la-Torre[1,2], Eric Granger[1], Robert Sabourin[1], Dmitry O. Gorodnichy[3]

[1] Laboratoire d'imagerie de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montréal, Canada

[2] Centro Universitario de Los Valles, Universidad de Guadalajara, Ameca, México

[3] Science and Engineering Directorate, Canada Border Services Agency, Ottawa, Canada

Paper accepted with revision in the journal "Machine Vision and Applications" from Springer, August 2014

ABSTRACT

Recognizing individuals of interest from faces captured with video cameras raises several challenges linked to changes in capture conditions (e.g., variation in illumination and pose). Moreover, in person re-identification applications the facial models needed for matching are typically designed a priori, with a limited amount of reference samples captured under constrained temporal and spatial conditions. Face tracking can however be used to regroup the system responses linked to a facial trajectory (facial captures from a person) for robust spatio-temporal recognition, and to update facial models over time using operational data. In this paper, an adaptive ensemble-based system is proposed for spatio-temporal face recognition (FR). Given a diverse set of facial captures linked to a trajectory of a target individual, an ensembles of 2-class classifiers is designed. A pool of ARTMAP classifiers is generated using a dynamic PSO-based learning strategy, and classifiers are selected and combined using Boolean combination. To train classifiers, target samples are combined with a set of reference non-target samples selected from the cohort and universal models using One-Sided Selection. During operations, each individual-specific ensemble of the system seeks to detects target individuals, and may self-update their facial models, using facial trajectories. A learn-and-combine strategy is then employed to avoid knowledge corruption during self-update of ensembles, and a

memory management strategy based on Kullback-Leibler divergence allows to rank and select validation samples over time to bound the systems memory consumption. Spatio-temporal fusion is performed by accumulating classifier predictions over a time window, and a second threshold allows to self-update facial models. The proposed systems was validated in a passport checking scenario with real-world Face in Action videos that feature abrupt and gradual patterns of change. At the transaction level, results show that the proposed system allows to increase AUC accuracy by about 3% for scenarios with abrupt changes, and by about 5% with gradual changes. Subject-based analysis reveals the difficulties of face recognition with different poses, affecting more significantly the lamb- and goat-like individuals. Compared to reference spatio-temporal fusion approaches, results show that the proposed accumulation scheme produces the highest discrimination.

3.1 Introduction

In person re-identification (or search and retrieval) applications, target individuals of interest must be recognized from face images captured across a network of video cameras. Automated face recognition (FR) systems are increasingly employed for decision support in such applications (Best-Rowden *et al.*, 2013; Fischer *et al.*, 2011), where a human operator seeks to reliably detect the presence of target individuals in several video feeds. In this case, the facial regions captured in videos are matched against the facial models of target individuals (enrolled to the system). These facial models are usually designed with a limited amount face captures collected under controlled conditions. Each model may be defined as a set of one or more reference samples (for a template matching system), or a set of parameters estimated during training with reference samples (for neural or statistical classifiers). However, faces acquired under semi- and unconstrained capture conditions may match poorly with stored facial models, because operational environments are complex and change abruptly or gradually due to variations in pose, illumination, expression, etc.

Adaptive biometric systems may be used to update facial models over time, given a new block of reference data. Using adaptive ensembles has shown to provide a robust solution when lim-

ited data is available for system design, and to avoid knowledge corruption (De-la Torre *et al.*, 2012a; Pagano *et al.*, 2012; Polikar *et al.*, 2001). However, since the collection and analysis of new reference data (e.g., through re-enrollment) for system update is often expensive and not always possible, several adaptive techniques enable biometric systems to update with unlabeled operational data (Rattani, 2010; Roli *et al.*, 2007, 2008). In video surveillance applications, individuals may be tracked in a scene, and facial captures across multiple frames may be regrouped in facial trajectories,[1] integrating both time and space information (Matta and Dugelay, 2009). And accumulated matching scores or decisions for the facial captures in a trajectory allow for robust spatio-temporal recognition (Despiegel *et al.*, 2012; Ekenel *et al.*, 2010; Zhou *et al.*, 2004) and accurate self-update of facial models (De-la Torre *et al.*, 2014a; Franco *et al.*, 2010).

In a recent publication (De-la Torre *et al.*, 2014a), the authors proposed a general framework for partially supervised learning from facial trajectories, in which tracking and classification information are combined. When the face of a new person is first captured in a video, the tracker is initialized to follow that face across frames, and a facial trajectory is captured. Predictions from individual-specific ensembles of detectors (EoDs) are accumulated for a fixed size time window. A detection threshold is then estimated on validation trajectories, and applied to the accumulated predictions to provide overall decisions. When a new trajectory surpasses a second (higher) update threshold, the system performs self-update of the corresponding facial model using all facial captures linked to the high confidence trajectory. That framework also provides the mechanism to select non-target training samples, and to rank and select validation samples to be stored in a long term memory, permitting to limit memory consumption after each self-update. However, the thresholding scheme only allows adapting to gradual changes in the video surveillance environment (e.g., due to aging), and only the most recent validation trajectory is considered for threshold estimation, which degrades system's knowledge.

[1] A facial trajectory is defined as a set of facial captures (produced by face segmentation) that correspond to a same high quality track of an individual across consecutive frames.

In this chapter, an adaptive ensemble-based system is proposed for spatio-temporal FR in person re-identification (search and retrieval applications). This particular realisation of the system is inspired by the framework described in (De-la Torre *et al.*, 2014a), and considers not only gradual, but also abrupt changes in data, as found in real world person re-identification applications. When a new trajectory becomes available to design or update the facial model of an individual of interest, the target's facial captures are used to design an EoD. A pool of probabilistic Fuzzy ARTMAP 2-class classifiers (Lim and Harrison, 1995) is generated using a dynamic particle swarm optimization (DPSO)-based learning strategy (Connolly *et al.*, 2012), and is combined with previously trained classifiers. These base classifiers are selected and combined using Boolean combination (BC), which takes advantage of the ROC space to choose the desired operations point (Khreich *et al.*, 2010b). A learn-and-combine incremental learning strategy incorporates the new data from an update trajectory, yet avoids knowledge corruption during self-update of EoDs. To train 2-class classifiers, a variation of the One-Sided Selection (OSS) algorithm is employed to select non-target training samples, and avoid the bias through the non-target class. A data management strategy based on the Kullback-Leibler divergence (KL) allows to rank and select a fixed number of validation samples over time, and bound memory consumption. The system accumulates ensemble predictions in a fixed size time window, and an individual-specific detection threshold is applied for accurate spatio-temporal FR. If accumulated predictions surpass a second (higher) update threshold, the EoD will self-update the corresponding facial model with the input trajectory. Finally, decision and update thresholds for spatio-temporal fusion are re-estimated with accumulations from past and new update trajectories every time the system is self-updated.

Video sequences from the Carnegie Mellon University Face in Action (FIA) video FR database were used for validation. Videos were captured from 180 subjects with an array of 6 cameras over three sessions separated by a three-month interval. In this data, individuals were captured under semi-constrained conditions for a security check point scenario. When a sequence is presented to the system during operations, high confidence target trajectories are used for performance estimation and self-update. Three levels of evaluation are used for benchmarking

– transaction-based analysis (ROC and *precision-recall* spaces), subject-level analysis (Doddington zoo characterization), and trajectory-level analysis (overall system behavior over video sequences).

The rest of the chapter is organized as follows. Sections 3.2 and 3.3 provide a survey of techniques used for FR in video surveillance and adaptive biometrics, respectively. Then, the adaptive ensemble-based system proposed for spatio-temporal FR is presented. Section 3.5 describes the experimental methodology –protocol, data set and performance measures. Finally, results are presented and discussed in Section 3.6.

3.2 Video-to-Video Face Recognition in Person Re-identification

The search and retrieval of individuals previously seen over a network of cameras finds many applications in video surveillance. Person re-identification typically exploits clothing appearance and gait for short term re-identification (Satta, 2013) and/or classical biometric traits when clothing is not constant, e.g., for long term re-identification (Best-Rowden *et al.*, 2013; Fischer *et al.*, 2011). This chapter focuses in person re-identification based on facial captures from a network of video cameras.

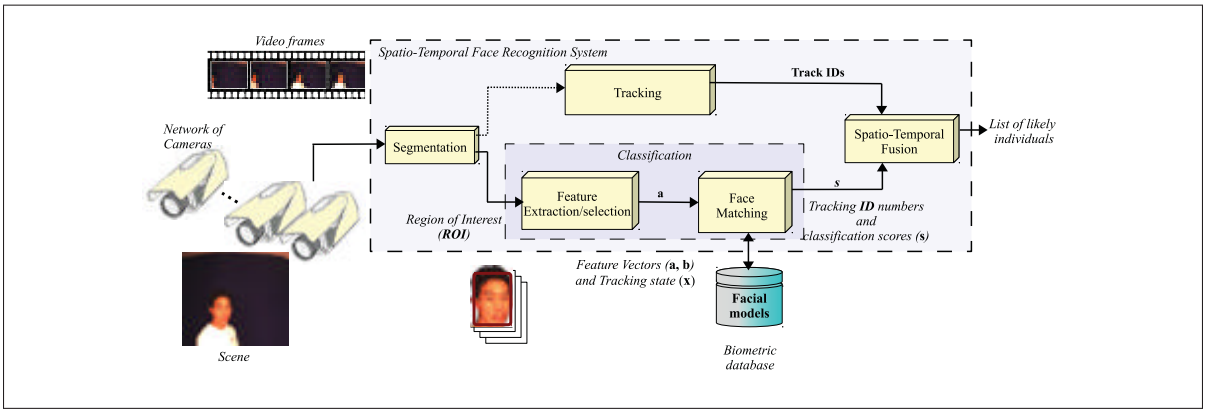


Figure 3.1 A generic track-and-classify system for spatio-temporal face recognition

Assume a system where 2D frames are captured from video streams using one or more IP cameras in a network. Spatio-temporal FR involves several processing steps, as shown in Fig. 3.1. First, the segmentation process isolates the facial regions of interest (ROIs) corresponding to faces captured in successive frames. Then, the feature extraction module extracts specific characteristics for tracking and classification. A tracker is typically initialized when a new person is viewed by a camera and emergent ROIs are detected far from other faces. A track is defined over consecutive frames using the state of the tracked facial region $\mathbf{x}$ (using appearance, scale, position, track number, etc.) and tracker-specific features (into vector $\mathbf{b}$). Invariant and discriminant classification features are extracted from each ROI (into vector $\mathbf{a}$): facial features can be categorized according to three levels of detail (Klare and Jain, 2010). Level 1 features contain low dimensional appearance information (e.g., principal component analysis on pixel intensities), level 2 features require information from the structure and specific shape and texture of the face (e.g., local binary patterns), and level 3 features are mostly used in forensic identification, and include scars, marks, and other micro features of the face. Tracking follows the movement or expression of distinct faces across video frames, whereas the classification function compares the ROIs to the facial models of individuals enrolled to the system. Finally, the decision function combines the tracking and classification information in order to predict a list of likely individuals in the scene.

Applications in video-surveillance include still-to-video FR (e.g., watchlist screening) and video-to-video (e.g., person re-identification) FR. Although many systems neglect temporal information, and use video sequences as a source of isolated facial regions, it is possible to design facial models from video streams, integrating time and space information. In particular, the presentation order of the frames affects the recognition accuracy in spatio-temporal FR (Matta and Dugelay, 2009).

Two variants can be distinguished among spatio-temporal FR approaches. *Tracking-then-recognition* approaches use segmentation to first crop a detected face, and then track the facial region over time. These approaches typically perform face matching on each frame, and then use majority voting for a final result. *Tracking-and-recognition* approaches attempt to simul-

taneously track and recognize, and may combine temporal and spatial information in a unified manner (Barry and Granger, 2007; Ekenel *et al.*, 2010; Zhou *et al.*, 2004), or integrate tracking and recognition within a single algorithm (Franco *et al.*, 2010; Lee *et al.*, 2005; Matta and Dugelay, 2006). Table 3.1 shows a survey of approaches from each category. In this chapter, spatio-temporal FR approaches with parallel tracking-and-recognition are considered for person re-identification.

Table 3.1 Categorization of approaches for FR in video in literature

Temporal Information		Approach
Neglect		Eigenfaces (Turk and Pentland, 1991) Fisherfaces (Matta and Dugelay, 2009) Active appearance models (Matta and Dugelay, 2009) Radial basis function neural networks (Matta and Dugelay, 2009) Elastic graph matching (Matta and Dugelay, 2009) Hierarchical discriminative regression trees (Matta and Dugelay, 2009) Unsupervised pairwise clustering techniques (Matta and Dugelay, 2009) Open Set TCM-kNN (Li and Wechsler, 2005) Ensembles of Fuzzy ARTMAP classifiers (Pagano *et al.*, 2012)
Exploit	Tracking-then-recognition	Fisherfaces with facial optical flow (Chen *et al.*, 2001) Dictionary-based face recognition (Chen *et al.*, 2014) Score and quality driven matching (Despiegel *et al.*, 2012) HMM extension for video (Liu and Cheng, 2003)
	Tracking-and-recognition	What-and-Where fusion Neural Network (Barry and Granger, 2007) Local appearance-based face models (Ekenel *et al.*, 2010) Tracking and Recognition using Probabilistic Appearance Manifolds (Lee *et al.*, 2003, 2005) Stochastic tracking and recognition through particle filtering (Zhou *et al.*, 2004) GMMs on unconstrained head motion (Matta and Dugelay, 2006) Recognition confidence and interframe continuity (Franco *et al.*, 2010)

3.2.1 Face Tracking

Facial tracking (FT) techniques allow to follow the location of each of individual and to regroup facial regions of a same person (without knowing his identity). The input of the tracker is the stream of frames coming from a video camera, and the initial face ROIs to be tracked, while the output track ID and defines a trajectory (set of ROIs with the same **ID**) for which the track maintains a high tracking quality Q_T. As a result, facial regions are regrouped as belonging to the same individual. Note that only the first ROI in a trajectory (ROIs from segmentation,

used for classification) has an exact match in a track (state of facial regions from the tracker) (Yilmaz *et al.*, 2006).

The basic tracking steps are face representation, prediction filtering and data association. In face representation, the tracked facial region is represented with distinctive features that permit to track the face from one frame to the next. Commonly used features are color histogram, skin color probability map, Eigenfaces and active contours, just to mention a few. Predicting the next state with Kalman and Particle filters seeks the new state **x** (appearance, scale, location, and/or velocity, etc.) of the facial region to be tracked in the current frame, based on the information in the previous frames and some underlying model for state transitions. The objective of the prediction filtering is to avoid drift and reduce the search space by using a probability framework, although some methods perform data association heuristically instead (e.g. Mean-shift and CAM-shift). Finally, in the data association step, the tracker associates a feature vector of the facial region extracted from the previous frame with the feature vector in the current frame. Tracking methods are often categorized according to the type of descriptor used for face representation: holistic, contour-based, and hybrid information. Most face-tracking methods in literature rely on holistic representations due to their robustness (Dewan *et al.*, 2013).

3.2.2 Face Matching

FR systems used for person re-identification usually consider an open set problem, with the premise that the number of individuals of interest is greatly outnumbered by non-target individuals. A multi-class classifier designed to reject unknown individuals in video FR is the Open Set TCM-kNN (Transduction Confidence Machine-k Nearest Neighbors) proposed by Li and Wechsler (Li and Wechsler, 2005). It provides a local estimation of the likelihood ratio used for detection, based in the relation between transduction and Kolmogorov complexity. The rejection threshold for never enrolled individuals is based on the distribution of the peak-side-ratio that characterizes the distribution of p-values that approximate the randomness deficiency. The p-values are constructed using the strangeness measure, which is the ratio of the sum of the k

nearest distances from the same class, divided by the sum of the k nearest distances from all other classes (Li and Wechsler, 2005).

Similarly, modular architectures with one detector per individual have been proposed, using 1- or 2-class classifiers per individual of interest. The advantages of these approaches has been widely studied in biometrics literature, and include the convenience for enrolling individuals and optimizing individual-specific parameters (Jain and Ross, 2002; Pagano *et al.*, 2012). For instance, Kamgar and Parsi propose an approach based on the identification of the decision region(s) in the feature space of individual-specific faces by training a dedicated feed-forward neural network for each individual of interest (Kamgar-Parsi *et al.*, 2011). Tax and Duin proposed a heuristic to combine any type of one-class classifiers for multi-class classification with outlier rejection. It allows to adjust the rejection threshold per individual, and to combine models that are not based on probability densities. By doing this, they combine classification scores from different probability densities for accurate FR (Tax and Duin, 2008). Another example is the SVM-based modular system proposed by Ekenel et al., applied to a visitor interface scenario (Ekenel *et al.*, 2010).

Finally, given the limited reference samples and the complexity of environments, modular approaches have been extended to train an ensemble of classifiers per individual. An ensemble of detectors (1 or 2-class classifiers) may be designed for each individual in a watch list. For classifier design, non-target samples are retrieved from the cohort model (CM, database maintained with trajectories from non-target individuals of interest) and the universal model (UM, database with training samples from unknown people appearing in scene. For example, Pagano et al. (Pagano *et al.*, 2012) proposed ensembles of 2-class classifiers co-jointly trained using a DPSO based training strategy. It allows for the generation of a diversified pool of ARTMAP classifiers that are selected and combined in the ROC space using Boolean combination (BC) (Pagano *et al.*, 2012).

3.2.3 Spatio-Temporal Fusion

Spatio-temporal FR approaches merge spatial information (e.g. face appearance) with the sequential variations presented over time (e.g behavior). Many of these approaches internally implement a tracking-like algorithm, whereas others take advantage of mature state of the art trackers to build trajectories. Regardless, motion information and matching scores or decisions are combined over time, and the matching performance may be improved on a time scale that is larger than the frame rate. As the tracker follows a face in the scene, it defines a track with all the followed regions, and a trajectory is defined as a set of facial ROIs (produced by face segmentation) that correspond to the same quality track of an individual across consecutive frames.

Liu and Chen used HMMs to model the appearance and dynamics of a person, obtaining high confident results on sequences that were then used to adapt the facial models. This approach merge spatial and temporal information within the HMM by modeling the probability distributions of the motion, and select the highest likelihood score provided by the HMM to decide the identity of the test video sequence. Authors compare their approach to a baseline system that performs IPCA recognition and apply majority voting for the identification decisions over the whole sequence (Liu and Cheng, 2003).

Probabilistic appearance manifolds expressed as a collection of subsets (pose manifolds) were used in video-based face recognition. In this approach exemplars are sampled from videos, and clustered with K-means, learning the probability between pose manifolds from training videos (Lee *et al.*, 2003). Zhang and Martinez divide the facial ROI in several sub-regions, and use an estimation of optical flow to weight the importance of each of them when estimating posterior probabilities. This technique allows to consider the motion between each pair of frames, including information from changes of expression (Zhang and Martınez, 2006).

Evidence accumulation strategies have shown to be gaining more interest in the field. They take into account multiple consecutive frames and allow to integrate matching responses over time. In the framework for video FR proposed by Gorodnichy in (Gorodnichy, 2005), different

strategies are mentioned for the sequential combination of output scores (postsynaptic potentials, or PSP) obtained from video frames. These combinations include (1) applying a threshold to the output scores of a sequence, (2) average or median of several consecutive frame decisions, (3) average or median of several consecutive PSP outputs, and (4) any combination of the above.

In the approach proposed by Zhou et al., the movement and identity are characterized using a motion vector and an identity variable (Zhou *et al.*, 2004). They estimate the joint posterior distribution of the motion vector and identity variable by combining three equations. The identity equation that governs the temporal evolution of the identity variable is given by

$$\theta_t = f(\theta_{t-1}, u_t); \text{ for } t \geq 1 \tag{3.1}$$

where u_t is a noise model, a common selection of θ_t is the affine motion parameters, and a common choice of the f function is an additive function. The motion equation governs the behavior of the tracking motion vector assume that the identity does not change over time:

$$n_t = n_{t-1}; \text{ for } t \geq 1 \tag{3.2}$$

The observation equation establishes the link between the equations 3.1 and 3.2, and is given by

$$z_t = \mathscr{T}\{y_t; \theta_t\} = g_{n_t} + v_t; \text{ for } t \geq 1 \tag{3.3}$$

where v_t is the observation noise at time t, and $\mathscr{T}\{y_t; \theta_t\}$ is a transformed version of the observation y_t. Then, the overall state transition probability is given by

$$p(x_t|x_{t-1}) = p(n_t|n_{t-1})p(\theta_t|\theta_{t-1}) \tag{3.4}$$

The What-and-where fusion neural network was applied for video-to-video FR of individuals in Video Surveillance (Barry and Granger, 2007). In this fusion scheme, an evidence accumulation module accumulates the classifier responses according to each track. The predictions of

this network are the results of multiple responses by the classifier, and in the particular implementation given in (Barry and Granger, 2007), the evidences are accumulated at in the category choice function T_h of a fuzzy ARTMAP neural network classifier. The evidence accumulation field F_h^e is connected to a track h, and its output prediction is given by

$$K^e = arg \max_{k^e} \{T^e_{hk^e} : k^e = 1,2,...,L\} \tag{3.5}$$

where L is the number of output class nodes.

Ekenel et al. (Ekenel *et al.*, 2010) present another example is the video-to-video FR system that progressively combines scores of the matchers using a sum rule over the full sequences to estimate the identity in video. In this approach, classification is performed using k-NN on a DCT representation of face images, and a min-max normalization is applied to the distance-based output scores. Weighted sum variants were also proposed and analyzed using the distance-to-model, distance-to-second-closest and a combination of both. The three frame weighting schemes allow to implement a more sophisticated spatio-temporal weighted sum. The distance-to-model (DTM) weights are given by

$$w_{DTM}(f_i) = \begin{cases} 1 & \text{if } d(f_i,c) < \mu \\ e^{-\frac{d(f_i,c)-\mu}{2\sigma^2}} & \text{otherwise} \end{cases} \tag{3.6}$$

where $d(f_i, c_{f_i})$ is the distance of all frames to the closest representative class c_f, i is the frame counter, and μ and σ are the mean and variance of the distribution of frame distances, estimated on an independent set. The the distance-to-second-closest weighting scheme is given by

$$W_{DT2ND}(f_i) = \varepsilon(\Delta(f_i)) = 1 - e^{-\Delta(f_i)} \tag{3.7}$$

where $\varepsilon(x,\lambda) = 0.1\lambda e^{-\lambda x}$, with $\lambda = 0.5$ is the distribution of frame distances to the second closest, and $\Delta(f_i)$ is the difference of distances to the closest and second closest. The frame-wise fusion scheme employs the sum-rule over all the sequence, adding the scores for all the ROIs in each trajectory T:

$$ss\_Decision(T) = \sum_{ROI_i \in T} score(ROI_i) \tag{3.8}$$

where $score(ROI)$ is the matching score that may be produced with any of their proposed weighting schemes.

Two methods were analyzed by Despiegel et al. (Despiegel *et al.*, 2012) to summarize the processing of video images from video sequences for a border control system. In the *score driven* method, facial regions are continuously matched against facial models until a matching score is above a predetermined threshold, which indicates a positive identification of the sequence. This method considers the highest matching score given by

$$ms\_Decision(T) = \max_{ROI_i \in T} \{score(ROI)\}, \tag{3.9}$$

and apply a predefined decision threshold. In the *quality driven* method, images are processed until a quality intrinsic to the considered image is above a predefined threshold, and the matching score over the predetermined threshold indicates a positive identification of the sequence. They observe in their experiments that when using score driven methods, the operational FPR cannot be computed off line. And using quality driven methods the off line DET curve corresponds to operational performances.

A dictionary-based method was proposed for person recognition in unconstrained environments, which builds video-dictionaries for still images to encode temporal, pose and illumination information (Chen *et al.*, 2014). This method takes advantage of the face and body traits, and apply kernel methods to learn nonlinearities to design several sub-dictionaries that encode distinct captures of the traits into biometric models. The minimum residual $\mathscr{R}$ for a ROI pattern indicates that the ROI is closest to one of the sub-dictionaries represented in the feature space, and is closely related to distance-based scores. They use majority vote for sequence-level decisions for identification, and minimum residual among all images in the sequence for verification.

$$mv\_Decision(T) = \arg\max_{\omega=\Omega} \left\{ \sum_{ROI_i \in T} d_i^{\omega} \right\}, \tag{3.10}$$

where d_i^{ω} is the classification decision corresponding to the ROI_i given the class $\omega \in \Omega$. The classes in Ω are the labels assigned to the individuals enrolled to the system. In mathematical notation, d_i^{ω} is given by

$$d_i^{\omega} = \begin{cases} 1 & \text{if the classifier decides class } \omega \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

Franco et al. (Franco *et al.*, 2010) propose a system that exploits recognition confidence (RC) and interframe continuity (IC) for template update. Although this system evaluates the recognition rate as the raw percentage of correctly recognized ROIs in a closed-set scenario (no time information), it uses the IC to decide if a ROI is suitable to be added to the gallery. The RC condition for a ROI pattern $\mathbf{a}$ given the two templates Υ_{i1} and $Upsilon_{i2}$ closest to $\mathbf{a}$, is fulfilled if

$$\frac{d(\mathbf{a}, \Upsilon_{i2}) - d(\mathbf{a}, \Upsilon_{i1})}{d(\mathbf{a}, \Upsilon_{i1}} > RC_{thr}, \tag{3.12}$$

where $d(\mathbf{a}, \Upsilon)$ is the distance between the feature vector $\mathbf{a}$ and a template Υ. The individual-specific threshold RC_{thr} is pre-fixed and updated according to the frequency of detection of the subjects. The IC condition is fulfilled if

$$\exists (i_1, v', p', s') \in X | (||p - p'|| < IC_{thr}) \wedge (v' \geq v_{thr}), \tag{3.13}$$

where i_1 is the identity according to previous detections, and v' is the amount of times a face was detected close to p' with scale close to s'. X is the set with new candidate faces, p and s are the position and scale corresponding to the previous candidate face, and IC_{thr} and v_{thr} are thresholds estimated according to the ITU algorithm (Franco *et al.*, 2010).

Finally, a framework was recently proposed for the combination of responses produced by commercial-of-the-shelf systems that for still-to-still FR over multiple frames (Best-Rowden

et al., 2013). Two fusion levels were distinguished, which allow for different combination schemes. Rank-level fusion allows to use combination schemes like majority voting to produce a single decision based on the responses produced for all frames in a trajectory. On the other hand, score-level fusion allows combination schemes like min, max, medium or averaging rules to reduce the theoretical error and improve empirical performance over using single-sample decisions.

3.2.4 Key Challenges in Person Re-Identification

One of the main challenges is that facial models are designed a priori with a limited number of reference facial captures. Facial models so designed are also limited in their representativeness, and yield poor accuracy when matched against faces captured during operations, under semi or uncontrolled capture conditions, and possibly with different cameras. However, evidence accumulation in spatio-temporal approaches allows to increase the overall matching performance of FR systems by using several captures in the final prediction.

Another challenge is the representativeness of facial models over time. Facial captures incorporate considerable variations due to limited control over operational conditions when images are acquired from unconstrained scenes (e.g., illumination, pose, facial expression, orientation and occlusion). New information may suddenly emerge during operations, and previously acquired data may eventually become obsolete in changing environments. Moreover, the physiology of the individuals may change over time, either temporarily (e.g., haircut, glasses, etc.) or permanently (e.g., scars, aging). These factors result in facial models that diverge over time with respect to the underlying data distributions. Automatic FR systems capable of adapting facial models over time constitute a potential solution to maintain or improve performance.

Facial models designed with a few frontal facial regions captured under controlled conditions are not expected to provide a high level of performance when matched against faces captured in different conditions with changes in illumination, pose, aging, etc. High quality face tracks allow to regroup ROIs that correspond to the same individual. Thus, if a facial model is updated

with a facial trajectory, it may be enriched with a variety of information that may not be possible to be automatically acquired from a single ROI.

In this chapter, a specialized system is proposed for video-to-video FR in person re-identification. It is composed of adaptive individual-specific EoDs, and uses evidence accumulation with a fixed size window over trajectories for robust spatio-temporal recognition.

3.3 Update of Facial Models

Several approaches allow for supervised update, providing reliable results (Connolly *et al.*, 2012; De-la Torre *et al.*, 2012a; Tax and Duin, 2008). However, obtaining new labeled reference data is often costly or impractical. To overcome this difficulty, several *semi-supervised* methods have been introduced for automatic template update during operational phases (Franco *et al.*, 2010; Okada *et al.*, 2001; Rattani *et al.*, 2009b, 2008a; Roli *et al.*, 2007, 2008; Roli and Marcialis, 2006). This chapter focuses on using self-updating algorithms to adapt facial models of a video-to-video FR system using trajectories.

3.3.1 Adaptive Biometrics

Self-update techniques (Roli *et al.*, 2007; Roli and Marcialis, 2006) were proposed to update biometric models based on the classification score produced by the system given an input biometric sample. The system is initially designed using reference samples from a set D_L of labeled data, and a set of unlabeled data D_u is employed for semi-supervised learning. A decision threshold γ^d is applied to the similarity scores generated after matching unlabeled samples. Then, samples with scores that surpass a higher updating threshold, $\gamma^u \geq \gamma^d$ (i.e., matched with a high degree of confidence), are used to update the corresponding biometric model. The subjective notion of *high degree of confidence* depends on both the application domain and matching algorithm, and usually the update threshold is chosen to be higher or equal than the decision threshold.

The advantages of adapting a biometric system using operational data carries an inherent risk. There exists a trade-off between the false updates and false rejections that affect of performance. A conservative threshold (or other parameters in the biometric model) may allow a system without false updates, but also a system that is never adapted to changes in the environment. Conversely, a less conservative threshold may lead to an increase in the number of false updates and the inherent deterioration of biometric models. An accurate selection of adaptation criteria (decision and adaptation thresholds) is crucial in the design of such systems.

Another technique that is commonly used in semi-supervised learning is called co-update. This strategy is adapted for use with two diversified matchers with independent galleries specialized on different biometric traits, modalities or scores, designed to mutually improve performance. The biometric traits originally used are the fingerprints and the face, where co-training is used to update the template-based face and fingerprint models (Roli *et al.*, 2007).

Different semi-supervised approaches have been proposed in literature, where statistical or neural classifiers are used to design the biometric models. For instance, a view representation that combines facial and torso-color histograms was used with bunch graph matching for adaptive person recognition (Okada *et al.*, 2001). This system is able to update existing biometric models, and automatically enroll unknown individuals based on a double thresholding strategy. Update is performed on operational video streams that provide high sequence-to-entry similarity, measure of confidence. The sequence-to-entry similarity is the average of maximum frame-to-entry similarity values, which in turn was defined as the maximum similarity value over all facial representations in a database entry (Okada *et al.*, 2001). Bayesian networks were also used for facial expression recognition and face detection using a stochastic structure search algorithm (Cohen *et al.*, 2004). This approach combined labeled and unlabeled data to train the classifier and search for the Bayesian network structure that provided the minimum probability of error, using maximum likelihood estimation. SVMs with locality preserving projections have also been combined to update facial models, by incorporating information from operational ROIs taken from video (Lu *et al.*, 2010). The algorithm first builds a data model

of a video sequence, and then uses semi-supervised locality preserving projections to build a graph with the geometrical structure of the face space.

MCSs have also been used in conjunction with the co-training and self-training. For instance, Didaci and Roli (Didaci and Roli, 2006) proposed an ensemble of five classifiers was trained with two different diversity generation techniques (bootstrap and the training of different classifiers). These techniques are based on a re-training schema for biometric model updates, and improve accuracy by 18% on the test set composed of the training (labeled) and unlabeled data, using the product rule for combination. In another variation, the co-training algorithm for MCS was proposed for updating only unlabeled samples that produced high confidence (El Gayar *et al.*, 2006). The five patterns with highest probability of belonging to the specific person, were selected as the most confident. This system was tested with 3 non-homogeneous classifiers in the ensemble, and provided the highest performance with a voting combination scheme. Finally, a semi-supervised classification schema based on random subspace dimensionality reduction was proposed for graph-based semi-supervised learning. In this approach, a kNN graph is built in each processed random subspace, and semi-supervised classifiers are trained on the resulting graphs, using majority voting rule for combination (Yu *et al.*, 2012).

MCSs for semi-supervised learning in the literature have provided improved accuracy, and show the utility of unlabeled samples. In this chapter, an adaptive MCS is proposed for video-to-video FR, that allows for semi-supervised learning from facial trajectories. It exploits the two thresholds (γ^d and γ^u) for self-update, and the quality of tracking as a second source of confidence, a characteristic borrowed from the co-update algorithm. The tracking quality allows to regroup facial regions from the same individual, and the accumulation of the positive predictions of each individual-specific ensemble over time allow for high confident decisions.

3.3.2 Adaptive Face Recognition Systems

Adaptive FR systems in literature have traditionally incorporated newly-acquired reference samples to update the selection of a user's template from a gallery, via clustering and editing

techniques. These systems allow to augment the representation of the intra-class variations in facial models.

Recent work on supervised incremental learning of facial models includes a FR system that relies on an adaptive MCS. An incremental learning strategy based on DPSO has been proposed to update an ensemble of incremental learning classifiers based on new data for video-based access control. It allows the evolution of an ensemble of heterogeneous multi-class ARTMAP classifiers from new reference data, using an LTM to store validation samples for fitness estimation and determining the number of training epochs. This approach reduces the effect of knowledge corruption by integrating information from diverse classifiers that are guided by a population-based evolutionary optimization algorithm (Connolly *et al.*, 2012). Another adaptive MCS that allows for design and update of facial models is composed of an ensemble of binary detectors (EoDs) per individual, an LTM and a dynamic optimization module. When a new data block becomes available, a diversified pool of 2-class ARTMAP classifiers is generated using a learning strategy based on the DPSO optimization algorithm. The combination function is updated using Boolean combination (BC) (De-la Torre *et al.*, 2012a). Learn++ is another well-known ensemble-based technique for incremental learning that has been tested on FR problems. This technique was proposed by Polikar et al. (Polikar *et al.*, 2001), and is inspired by the AdaBoost algorithm. It allows for supervised incremental learning by incorporating a new set of classifiers to the ensemble each time new data becomes available. The generation of weak classifiers is performed using a bagging strategy, by training distinct base classifiers on bootstrap replicates of the training set.

Semi-supervised approaches for facial model update are generally based on the classification similarity. For instance, in (Roli and Marcialis, 2006), self-training has been applied to a FR system using Euclidean distance. In each iteration, the PCA-based feature space is updated with the newly acquired soft-labeled samples. In (Hewitt and Belongie, 2006), a method is proposed to combine tracking and recognition to build facial models based on co-training. This method is used to label facial samples, and thus to build a learning dataset for each user. Their initial facial model consists of a single manually selected frontal image, and the extrac-

tion of new face samples is done off-line. They use a tracker instead of a second classifier to identify informative training examples. The Graph Mincut method, which can be seen as self-update given that the system updates its own templates using the recognition scores thrown by the same templates, has been proposed to update templates by analyzing the underlying structure of input operational data (Rattani *et al.*, 2008a). In this extension, a pair-wise similarity measure between operational face captures and existing templates is used to draw a graph that relates these samples, allowing for a global template optimization. A system that exploits classification similarity and video information, is presented in (Franco *et al.*, 2010), to perform incremental template update. It is based on the similarity between acquired facial images and existing templates, and exploits the frequency of detection on the complete sequences of the different subjects in the scene. Recognition confidence and interframe continuity measures were integrated in a face recognition system that can assign unlabeled images to subjects in the gallery. When these two quantities surpass independent individual-specific thresholds, templates are recognized as belonging to a subject, and are incorporated to the gallery.

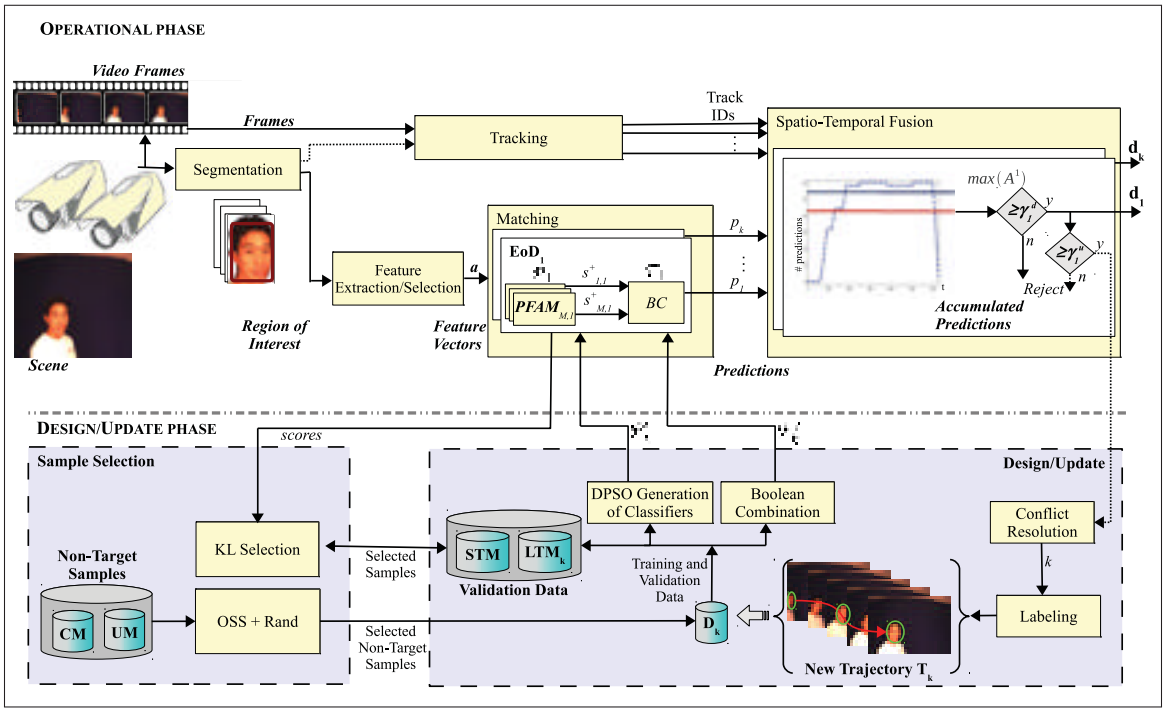


Figure 3.2 Block diagram of the proposed adaptive spatio-temporal system for video-to-video FR

Finally, a framework that allows to combine tracking and classification information to recognize individuals in video-to-video FR was proposed in (De-la Torre *et al.*, 2014a). In that framework, decisions from classifiers corresponding to consecutive ROIs in a trajectory are accumulated over time, allowing to estimate a decision threshold γ_k^d based on the maximum difference between the positive accumulation curve and the higher negative envelope. An update threshold $\gamma_k^u > \gamma_k^d$ is applied to the accumulated decisions of a trajectory in order to decide if that trajectory can be used for self-update. In that way, the system can incorporate new knowledge from high confident trajectories acquired during operations. The next section describes a particular realization of a self-updating system for spatio-temporal face recognition, based on the aforementioned framework.

3.4 A Self-Updating System for Spatio-Temporal Face Recognition

The structure of the adaptive MCS for video-to-video FR is shown in Fig. 3.2. It is composed of 7 subsystems: 5 used in normal operation and 2 used in the design/self-update phase. The segmentation module is used for face detection, the feature extraction/selection module and the matcher with one EoD per enrolled individual produces classification predictions. The IVT face tracker follows faces in scene allowing the spatio-temporal fusion system to regroup and accumulate target predictions over a fixed size window for enhanced spatio-temporal FR. Detection (γ_k^d) and update (γ_k^u) thresholds for spatio-temporal fusion are estimated using validation trajectories, and the design/update module avoids knowledge corruption by using a learn-and-combine strategy. Individual-specific EoDs are designed by the design/update module, by training a pool of PFAM 2-class classifiers using a DPSO training strategy, and estimating the fusion function with BC. The sample selection system allows to reduce the negative bias of the training and validation sets using the OSS and random selection strategies.

In the *operational phase*, the feature vectors corresponding to ROIs captured in scene are matched against facial models. In the matching process, the scores produced by the PFAM classifiers in the EoDs are thresholded and combined with the operations point selected on validation with BC. Target predictions for a trajectory T produced by the EoDs are then ac-

cumulated over a fixed size time window, and the detection γ_k^d and update γ_k^u thresholds are applied to the resulting accumulation. When the accumulation of predictions from the EoD_k surpasses the detection threshold γ_k^d, the individual k is positively detected. If it surpasses the update threshold γ_k^u, the trajectory T is assigned to individual k (T_k), and the self-update process is triggered. The impact of different sizes of the time window, as well as different detection and update thresholds are analyzed in ROC space (see Section 3.6).

The *design/update phase* starts when a labeled (or self-labeled) trajectory T_k becomes available. The sample selection subsystem allows to build a labeled training set D with all target ROIs from T_k, and a combination of non-target samples selected from the CM and UM. The proposed OSS+Rand selection combines the target samples with non-target samples in the borderline between distributions, and samples that are representative of the non-target data distribution. The design/update subsystem splits D into learning and validation subsets that are used to generate a new pool of PFAM classifiers, which in turn, is integrated with the old pool. Then, a mix of new and old validation samples from the LTM is temporarily stored in the short term memory (STM). BC is used to select and combine the classifiers in the pool using a subset of the validation samples in the STM, and the remaining samples are used to select the operations point (e.g. at $fpr = 1\%$). Finally, the validation samples in the STM are ranked and selected using the KL divergence, and the most relevant are stored in the LTM for further validation.

3.4.1 Modular Classification System

The modular classification system is composed of individual-specific EoD that allow for enhanced classification accuracy when only a limited amount of data is available for system design (De-la Torre *et al.*, 2012a; Pagano *et al.*, 2012). Accordingly, each EoD estimates discriminant bounds between the target (individuals of interest) and non-target (the rest of the world) populations. Each ensemble EoD_k is comprised of a pool of Probabilistic Fuzzy ARTMAP (PFAM) classifiers (Lim and Harrison, 1995) $\mathscr{P}_k = \{c_{1,k}, ..., c_{M,k}\}$, and a fusion function $\mathscr{F}_k$ that is designed in the ROC space using the Boolean combination (BC) (Khreich *et al.*, 2010b).

The PFAM neural network classifier combines Fuzzy ARTMAP density estimation for learning category prototypes, with a non-parametric posterior probability distribution procedure inspired by the Probabilistic Neural Networks during the operational phase (Lim and Harrison, 1997). It allows for incremental learning of new data, and can learn quickly and efficiently with limited data. Additionally, the PFAM classifier allows to produce estimated posterior probabilities of class membership.

Given an input sample $\mathbf{a}$, the output prediction for each class $\omega_j \in \Omega$, $\Omega = \{\omega_1, \omega_2\}$, is represented by:

$$p(\mathbf{a}|\omega_j) = \frac{1}{(2\pi)^{M/2}\sigma_j^M} \exp\left\{ -\frac{(\mathbf{a} - \mathbf{w}_j^a)^T (\mathbf{a} - \mathbf{w}_j^a)}{2\sigma_j^2} \right\} \tag{3.14}$$

where σ_j is the variance represented by the ratio of the squared minimum distance between $\mathbf{a} - \mathbf{w}_j^a$ and any other center M-dimensional pattern, and $\mathbf{w}_j^a$ are the centers of mass corresponding to the category prototypes in the F_2^a layer inherited from the Fuzzy ARMAP architecture. The estimation of posterior probabilities for 2-class classifiers is given by:

$$\hat{P}(\omega_j|\mathbf{a}) = \frac{p(\mathbf{a}|\omega_j)P(\omega_j)}{\sum_{i=1}^{2} p(\mathbf{a}|\omega_i)P(\omega_i)} \tag{3.15}$$

where priors $P(\omega_j)$ are estimated based on the proportions of each class in the training data.

PFAM inherits four hyperparameters from its underlying Fuzzy ARTMAP architecture. These hyperparameters are the learning rate $\beta \in [0,1]$, the choice $\alpha > 0$, the match tracking $0 < \varepsilon \ll 1$, and the baseline vigilance $\bar{\rho} \in [0,1]$. A fifth hyperparameter r controls the overlap between probability densities for prediction using the probabilistic neural network.

An incremental learning strategy based on the DPSO algorithm allows to evolve a pool of classifiers in the five-dimensional space of hyperparameters $\mathbf{h} = [\alpha, \beta, \varepsilon, \bar{\rho}, r]$. It generates a diversified pool of PFAM classifiers taking advantage of the correlation between the diversity

within a dynamic particle swarm and the diversity within a corresponding pool (Connolly *et al.*, 2012).

Given a new training set D^t, and validation sets to stop training epochs (D^e) and for fitness evaluation (D^f), the algorithm initializes N PFAM networks and PFAM_n^{start}, and sets the swarm parameters with the initial iteration counter at $\tau = 0$. Then, the positions and velocities of the particles in the swarm are randomly initialized. The N particles (PFAM classifiers) are trained on D^t and D^e respectively, and the DPSO fitness function $f(\mathbf{h}_n(\tau), t)$ is the AUC of the ROC produced by the classifier after evaluation on D^f:

$$f(h_n(\tau)) = AUC(ROC(c_{\mathbf{h}_n(\tau)}, D^f)), \tag{3.16}$$

where $c_{\mathbf{h}_n(\tau)}$ is the classifier trained on D^t with the hyperparameter vector $\mathbf{h}_n$ at the iteration τ of the algorithm, and validated with D^e.

The iterative process starts by a random initialization of the N particle positions in the optimization space $\mathbf{h}$, and training the corresponding PFAM classifiers with $\mathbf{h}_n$ and a random pattern presentation order. The fitness for the N PFAM networks that correspond to each particle are evaluated, and those with highest fitness value are considered local bests. The old particles for which the fitness was improved are updated: the fitness, position and network associated with the PFAM networks are replaced. In the case that previous and new fitness is equal, the network with lower complexity (the least F_2 nodes) is chosen. The positions $\mathbf{h}_n$ are updated, and the procedure is repeated from the fitness evaluation. The process is repeated until the DPSO reaches the stopping condition, after fitness converges.

The fusion function $\mathscr{F}_k$ is estimated using Boolean combination (BC), and holds a set of operations points (maximum realizable ROC curve vertices). Thus, it provides an increased AUC that is equivalent or higher than the maximum realizable ROC (MRROC) of the ROC curves produced by the classifiers in $\mathscr{P}_k$. BC selects an ensemble from the pool of classifiers, and Boolean fusion functions and thresholds are adapted for improved accuracy. Initially, the algorithm receives the scores produced by the classifiers to be combined after presentation of the

combination set D^c, and starts by ordering the PFAM classifiers according to the AUC accuracy in decreasing order. All pairs of operations points in the ROC curves from the first two classifiers are combined using all Boolean functions, and the convex hull of the collection of original and new points is obtained. Then, the vertices in this convex hull are combined with the operations points in the ROC curve from the third classifier, and a new convex hull is obtained. The process is repeated until the ROC curves for all the classifiers are combined, and the convex hull that includes all the classifiers is obtained. The newly generated operations points are successively re-combined, in the same order, with operations points of the classifiers, until the overall convex hull stops improving (Khreich *et al.*, 2010b). The so estimated operations points –vertices of the final convex hull– are comprised of classifier specific thresholds and Boolean combination functions. The use of all Boolean functions in the combination allows this method to make no assumptions with respect to the independence of the classifiers. In practice, this technique has proven to be more accurate than majority voting, median (Khreich *et al.*, 2010a), and weighted majority voting (Learn++) (De-la Torre *et al.*, 2012b).

After the MRROC is estimated by BC, each vertex (operations point) is evaluated on an independent selection set D_k^s, which allows to select an unbiased operations point in the ROC space for a predefined fpr. If no operations point exists for the specified fpr, a virtual classifier is produced by interpolating the closest adjacent operating points (Fawcett, 2006).

During operations, the classifiers $c_{m,k}$ of each ensemble EoD_k, $k = 1,...,K$, $m = 1,...,M$, produces an output score $s_{m,k}^+(\mathbf{a})$ for a given ROI pattern $\mathbf{a}$. The scores are then combined using $\mathscr{F}_k$. Each individual-specific EoD_k produces an output prediction $p_k(\mathbf{a})$. Positive predictions are then accumulated over time for each trajectory in the spatio-temporal fusion system to produce a global decision (see Fig. 3.2). Finally, self-update is achieved by using adaptive ensembles of ARTMAP classifiers, each one capable of supervised and unsupervised incremental learning. A *learn-and-combine* strategy is employed to maintain performance even after several adaptations, yet avoid knowledge corruption associated with many incremental learning classifiers (De-la Torre *et al.*, 2012a).

3.4.2 Tracking System

The face tracker initializes a new trajectory with the first facial ROI captured by the segmentation system in a different area of the scene. Then, the individual is tracked independently. As the tracker follows the facial region through the scene, the segmentation system captures high quality facial ROIs for some of the frames, allowing to produce a trajectory. Note that the segmentation module does not retrieve a facial region from all frames. The diverse set of facial ROIs regrouped with the tracker belongs to the same individual. When the tracking quality Q_T for a trajectory T falls under a pre-defined quality threshold ($Q_T < \gamma^T$), the track is dropped, and its trajectory is closed.

The incremental visual tracker (IVT) is considered in the proposed system. Accurate data association is performed by updating a low-dimensional subspace that represents the appearance of each person's facial regions (Ross *et al.*, 2008). It adapts over time to changes in the appearance of the target face based on capture conditions. This generative method takes advantage of the Eigen-faces representation with particle filters, and data association is performed with Euclidean and Mahalanobis distances. Once a new person (or face) is initially detected, IVT uses template matching to track the face within the first n frames. Then, it defines a data block to compute an appearance-based face model represented in the Eigenspace spanned by these first n samples. The Sequential Karhunen-Loeve algorithm is used to update the Eigenspace and corresponding face representation.

The quality (confidence) of the new tracked face region measures the likelihood of that region to belong to the initial trajectory. In IVT, it can be derived from the observational model, given an image face patch $\mathbf{I}_t$ and a predicted position (particle) X_t, as

$$\begin{aligned} Q_T = p(\mathbf{I}_t|X_t) &= p_{d_t}(\mathbf{I}_t|X_t)p_{d_w}(\mathbf{I}_t|X_t) \\ &= \mathcal{N}(\mathbf{I}_t;\mu,UU^T+\varepsilon I)\mathcal{N}(\mathbf{I}_t;\mu,U\Sigma^{-2}U^T) \end{aligned} \tag{3.17}$$

where $p_{d_t}(I_t|X_t)$ is the probability of a sample generated from a subspace, and $p_{d_w}(I_t|X_t)$ is the likelihood of the projected sample within a subspace, modeled by the Mahalanobis distance

from the mean. I is an identity matrix, μ is the mean, and εI corresponds to the additive Gaussian noise in the observation.

3.4.3 Spatio-Temporal Fusion System

The adaptive MCS detects the presence of an individual of interest when on the number of positive predictions by EoD_k surpasses the detection threshold γ_k^d. Given a trajectory T, each EoD_k generates a prediction $p_k(\mathbf{a}_n)$ for each sample $\mathbf{a}_n$ associated with a $ROI_i \in T$. Output predictions from EoD_k over the ROI samples of a trajectory T, at the selected operations point, are defined by the set $\mathbf{P}_k = \{p_k(\mathbf{a}_1), ..., p_k(\mathbf{a}_N)\}$, associated with each input ROI sample $\mathbf{a}_n$. Negative predictions set $p_k(\mathbf{a}_n) = 0$, and positive ones set $p_k(\mathbf{a}_n) = 1$. The spatio-temporal fusion system accumulates the number of positive predictions A_k of each EoD_k on fixed size window W according to:

$$A_k = \sum_{i=0}^{W-1} p_k \cdot \mathbf{a}_{(W-i)} \quad \in [0, W] \tag{3.18}$$

For instance, a window of size $W = 30$ accumulates the last 30 predictions from the same trajectory. Each EoD_k accumulates a sequence of predictions that range from 0 (EoD_k made only negative predictions for W), to a maximum of W (EoD_k made only positive predictions for the last W ROIs).

Based on these accumulations A_k, for $k = 1, ..., K$, the system produces overall decisions. If A_k surpasses threshold γ_k^d, the system detects the presence of individual k and alerts the operator. Furthermore, if A_k surpasses the update threshold γ_k^u, the trajectory is used for self-updating of the corresponding EoD_k. Given the negative effects on performance caused by false updates, threshold γ_k^u is greater or equal to γ_k^d.

The detection threshold γ_k^d for each EoD_k is estimated using a validation set composed of one positive and several negative trajectories (see Fig. 3.3). In this way, a single target trajectory is required for design of the facial model. An accumulation curve is computed for each trajectory

within the validation dataset. The *higher negative envelope* (*hne*) is defined as the curve formed from the highest A_k values of the negative accumulation curves. The *positive accumulation curve* (*pac*) is the accumulated predictions over the trajectory for the corresponding individual k. The detection threshold for the EoD_k is computed by a weighted sum of two components, and is given by

$$\begin{aligned} \gamma_k^d &= w_1(\max\{pac(f_i) - hne(f_i) : i = 1,...,|T_k|\}) \\ &\quad + w_2(\max\{hne(f_i) : i = 1,...,|T_k|\}) \end{aligned} \tag{3.19}$$

where f_i is the frame number i in the given trajectory. The first term maximizes the capacity of the system to differentiate between target and non-target trajectories, and the second maximizes the correct rejection capacity. These weights remained equal and fixed in (De-la Torre *et al.*, 2014a), but an increase in the operational conditions eventually require increasing the importance for one term or the other. The weights must respect the constraint of $w_1 + w_2 = 1$ in order to avoid thresholds out of boundaries.

By considering the presentation order of the target and non-target ROI patterns, the time information is included in the threshold estimation for particular facial models. The adaptation threshold γ_k^u is set to a value equal to or greater than γ_k^d, and it is manually set according to prior knowledge:

$$\gamma_k^u = \gamma_k^d + \Gamma_k \tag{3.20}$$

where Γ_k is a user-defined real value between 0 and $(W - \gamma_k^d)$. Fig. 3.3 illustrates the measures used in the threshold estimation strategy, presenting the *pac* and the *hne*. The reliability of γ_k^d and γ_k^u estimates grows with the number of non-target trajectories present in the validation set.

When the accumulation from a trajectory T surpasses the detection threshold γ_k^d for one or more EoDs, the system outputs the corresponding decision signals. The output to the decision support system lists all individuals of interest that are detected in the scene. When the accumulation surpasses the update threshold γ_k^u, the corresponding trajectory is used for update of the classification system and the detection and update signals. The decision threshold is updated

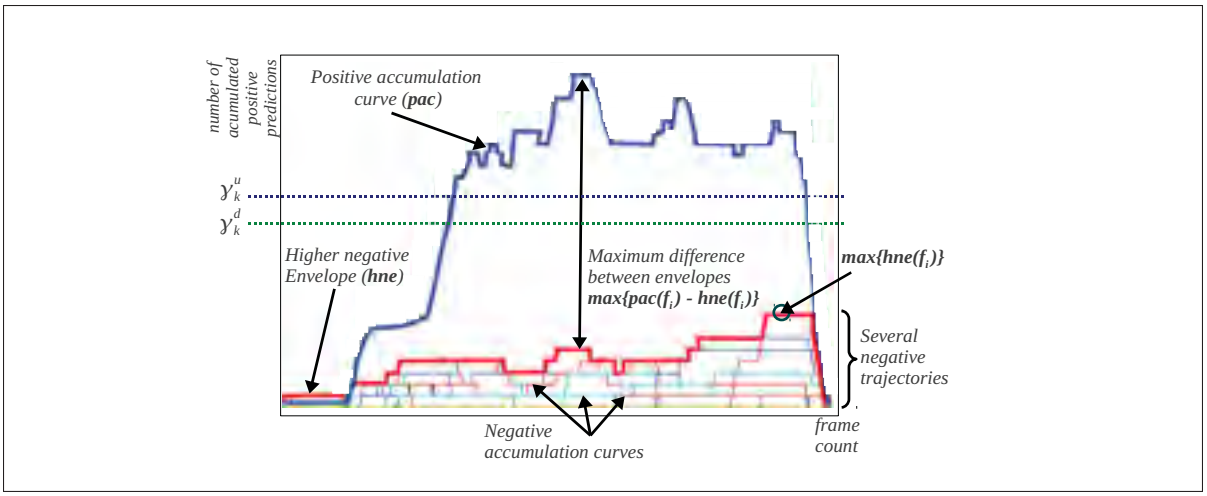


Figure 3.3 Estimation of detection and update threshold on validation trajectories at the decision level

considering the higher positive envelope (*hpe*) instead of the *pac* used for single trajectory scheme, as follows

$$\begin{aligned} \gamma_k^d &= w_1(\max\{hpe(f_i) - hne(f_i) : i = 1,...,|T_k|\}) \\ &\quad + w_2(\max\{hne(f_i) : i = 1,...,|T_k|\}) \end{aligned} \tag{3.21}$$

where the *hpe* corresponds to the highest accumulation values for all the overlapping target trajectories, assuming target trajectories start at the same point. This *hpe* represents the highest values obtained in the accumulation curves from past and current target trajectories, and define the highest possible values obtained for the target individual over time.

3.4.4 Design/Update System

When a new trajectory T_k is detected and labeled for design/update, all its facial ROIs from segmentation share the same label. Then, these facial ROIs are used to update the EoD_k, thus incorporating the diversified set of ROIs into the corresponding facial model. This greater diversity of samples is augmented when the captured ROIs over the trajectory present diversity of conditions (pose, lighting, etc). These samples allow for facial models that are more represen-

tative of video capture, increasing the capacity of the system to recognize faces from different capture conditions. In pattern recognition terms, the incorporation of these diversified set of reference samples allows to extend the boundaries between target and non-target individuals in the feature space, in accordance with the most recent facial ROIs.

Given a design/update trajectory T_k produced by the self-update system, the set with all its facial ROIs is divided into three subsets to follow a *learn-and-combine* strategy. Each combines target samples (from T_k) and non-target samples. An OSS based selection algorithm allows to retrieve borderline and distinctive non-target samples. It is used to select non-target samples from the CM and UM to populate the three training/validation data subsets (see Section 3.4.5). The CM database is comprised of a set of trajectories from other individuals of interest (excluding individual k); and the UM database is comprised of trajectories from other non-target individuals that represent the rest of the world, e.g. random individuals that appear frequently in the scene.

The first subset D^t is assigned for training[2], the second D^e for validation on the number of epochs that the classifiers are trained, and the third D_k^f for optimization of classifier hyperparameters. Then, the incremental learning strategy based on DPSO (Connolly *et al.*, 2012) is used to generate a diversified pool of classifiers, and add them to the previous pool $\mathscr{P}_k$. The validation sets (D^e and D^f) are then added to a short term memory (STM_k). At the first design step, the LTM_k is empty, however after the first adaptation, the validation samples in the STM_k are mixed with those stored in the previous LTM_k. Samples in both memories are combined, randomized and divided into two subsets. The first set D^c is used to select the classifiers from $\mathscr{P}_k$ to form the $\mathscr{F}_k$ and the EoD_k, and the second D^s to select the operations point in the ROC space. Given the modular architecture, the process is similar for samples stored for all the EoDs. In summary, an ensemble EoD_k is updated with new ROIs from a trajectory T_k by generating new base classifiers, adding these to a pool $\mathscr{P}_k$, and updating the fusion function according to the old and new validation samples.

[2]For simplicity of notation, in this chapter the k has been omitted from all design data blocks, e.g. $D_k^t \equiv D^t$.

Algorithm 3.1: Design and update of the EoD_k

Input : T_k, $EoD_k = \{\mathscr{P}_k, \mathscr{F}_k\}$, LTM_k, UM, CM
Output : EoD'_k, LTM'_k // $EoD'_k = \{\mathscr{P}'_k, \mathscr{F}'_k\}$ and LTM'_k
Divide T_k in D^t, D^e, D^f evenly // T_k (target samples)
$D^t \Leftarrow OSS\_NEG\_SEL(D^t, UM, CM)$ // 2-class sets
$D^e \Leftarrow OSS\_NEG\_SEL(D^e, UM, CM)$
$D^f \Leftarrow OSS\_NEG\_SEL(D^f, UM, CM)$
$P'_k \Leftarrow \{c'_{1,k}, ..., c'_{M,k}\}$ // Pool generated on D^t, D^e, D^f
$\mathscr{P}_k \Leftarrow \mathscr{P}'_k \cup \mathscr{P}_k$ // Combine old and new pools
$STM_k \Leftarrow D^e \cup D^f \cup LTM_k$ // Old and new samples
Divide STM_k in D^c and D^s evenly
$\mathscr{F}'_k \Leftarrow FUSION(D^c, D^s, fpr)$ // Fusion function
$EoD'_k \Leftarrow \{\mathscr{P}'_k, \mathscr{F}'_k\}$ // Updated EoD_k
$LTM'_k \Leftarrow KL\_SEL(STM_k, \lambda_k)$ // Use KL to manage LTM_k

Assuming that the size of the LTM_k for EoD_k is λ_k, the STM_k size is chosen to have at least $2\lambda_k$ in order to store enough new and old validation samples. Then, the validation samples in the STM_k are ranked according to Eq. 3.22 (see Section 3.4.5), and the λ_k samples with the highest values are stored in the LTM_k.

3.4.5 Sample Selection

Target samples from the design/update trajectory T_k are coupled with negatives from the CM and UM to form the learning set D. The OSS subsampling strategy (Kubat and Matwin, 1997) is employed to reduce the bias of training 2-class classifiers with imbalanced data sets (limited positive vs. abundant negative samples). This method preserves all target (minority class) samples and selects those non-target (majority class) samples that lie close to the area of overlap between classes. Then, those samples that are redundant, and those that are difficult to classify (involved in Tomek links) are discarded.

When a trajectory T_k is provided to the system for training/update, the corresponding ROIs are used to build dataset of positive samples D^+. A set of negative samples D^- is also built by subsampling from the UM and CM. The system applies the OSS algorithm to $D^+ \cup D^-$ to select a consistent subset for design of the binary base classifiers. The resulting dataset D comprises the complete set of positives D^+, as well as the negative samples selected by OSS

(close to the decision boundaries) $D_{oss}^{-'}$, and (3) a uniform random selection of negatives D_d^-. This algorithm makes no assumptions with respect to the probability distribution of the positive and negative samples. Border (selected by OSS) and non-border (randomly selected) samples are both included in D. The OSS algorithm permits an unbiased selection of negative samples, based solely on the distribution of the new samples.

Algorithm 3.2: *OSS_NEG_SEL.* Select non-target samples for system design

Input : D^+, UM // Samples from T_k, UM and CM
Output : D // Target and non-target samples
$D^- \Leftarrow UM \cup CM$ // All non-target samples
$[D_{oss}^+, D_{oss}^-] \Leftarrow OSS(D^+, D^-)$ // Select by OSS
$np \Leftarrow |D^+|$ // Number of target samples
$D_{oss}^{-'} \Leftarrow RAND\_SEL(D_{oss}^-, np)$ // Select np non-target
$D_d^- \Leftarrow RAND\_SEL(D^-, np)$ // Select np distinctive non-target from D^-, not selected by OSS
$D \Leftarrow D^+ \cup D_{oss}^{-'} \cup D_d^-$

Level **C** ranking measures permit the selection of samples from the LTM_k that are difficult to classify by the ensemble members. These samples are distinctive of the decision bound between the positive and negative classes, as estimated with the base classifiers. The disagreement of base classifiers on a determined validation sample is proportional to its difficulty, give a degree of information for border specification when the fusion function is estimated. This is also valid for the accurate selection of operations points. Among ranking measures available in the literature, only the Kullback-Leibler divergence produces a continuous measure of the disagreement between the ensemble members (De-la Torre *et al.*, 2013). The KL divergence of an input sample **a** is computed using:

$$KL(\mathbf{a}) = \frac{1}{M} \sum_{m=1}^{M} \left(\sum_{i \in \Omega} s_m^i(\mathbf{a}) \log \frac{s_m^i(\mathbf{a})}{\hat{P}_{EoD_k}^i(\mathbf{a})} \right) \tag{3.22}$$

where M is the number of classifiers in the ensemble EoD_k, and $\hat{P}_{EoD_k}^i(\mathbf{a})$ given by (3.23) is the consensus probability that the class $i \in \Omega$ is the correct label for sample **a**, given the scores $s_n^i(\mathbf{a})$ produced by the base classifiers:

$$\hat{P}^{i}_{EoD}(\mathbf{a}) = \frac{1}{M}\sum_{n=1}^{M} s^{i}_{n}(\mathbf{a}) \tag{3.23}$$

The value of KL divergence is proportional to the informativeness of a sample **a**. The most informative samples present the largest average difference between scores of any single committee member and the consensus.

Algorithm 3.3: LTM management using the KL div., $KL\_SEL(input = \{D, s_k(a_i), \lambda_k\}, output = \{Dr\})$

Input : D, $s_k(a_i)$, λ_k // Data block, scores $s_k(a_i)$, // $a_i \in D$ produced by EoD_k and // the size of the LTM_k

Output : Dr // Data block with λ_k representative samples from D

```
// For each sample in the data block
for a_i ∈ D do
    relevance_i = KL(s_k(a_i))   // Compute the KL divergence according to Eq. 3.22
D ⇐ SORT(D, relevance, dec) // Sort D in decreasing order, according to relevance_i
Dr+ ⇐ FIRST_POSITIVES(D, ⌈λ_k/2⌉)   // Positive samples with highest KL divergence
Dr− ⇐ FIRST_NEGATIVES(D, ⌈λ_k/2⌉)   // Negatives with highest KL divergence
Dr ⇐ Dr+ ∪ Dr−
```

Algorithm 3.3 details the procedure to select the most relevant validation samples from the whole validation set in the STM. Given an EoD, the *KL_SEL* algorithm allows for the selection of the most challenging samples from the validation set, providing information on the overlapping area according to the agreement of the ensemble members. When a validation dataset D is presented to the algorithm, all samples are ranked according to the KL divergence using the scores produced by base classifiers in the pool $\mathscr{P}_k$. The highest ranked samples are retained, while the less informative ones are discarded. Thus, the ranking method is based on past and present information on samples that are difficult to classify, according to older and newer classifiers.

Table 3.2 Parameters for all the blocks in the proposed adaptive system

Process	Technique	Parameter	Value
Face Segmentation	Viola-Jones	Pose and eyes training	Haar files from OpenCV
		Scale factor	1.1
		Minimum overlapping detections	2
		Flags	Scale Image
		Smallest region	0.1×0.1 the size of the image
Face Tracking	Incremental Visual Tracker	Particle filters	Standard (Ross *et al.*, 2008)
		Batch size	5
		Forgetting factor	0.9
Feature Extraction	MLBP	Block sizes	3×3, 5×5, 9×9
	PCA	Principal components	32
Learn-and-combine	DPSO	Initial particles in swarm	60
		Particles per sub-swarm	5
		Sub-swarms	6
		Maximum iterations	30
		Extra iterations	5
	IBC	Iterations	1
		Points in ROC curves	All
	LTM management	λ_k	100
Decision fusion	Window accumulation	Window size	30 frames
		Weights	$w_1 = w_2 = 0.5$
		Update discrimination	$\Gamma_k = 1$

3.5 Experimental Methodology

The experiments described in this section follow a common evaluation protocol for adaptive systems, which divides the design-update data into different subsets, and a separate independent test set that represents the reference never seen operational environment (Singh *et al.*, 2010; Roli and Marcialis, 2006; Franco *et al.*, 2010; Liu and Cheng, 2003). The main goal in these experiments is to characterize the proposed adaptive video-to-video face re-identification system in two semi-constrained environments, considering different video sequences that present changes in age and pose for the same individual.

The parameters for each block belonging to the system are summarized in Table 3.2. For face segmentation, tracking and feature extraction, the standard parameters were used according to the published references. The parameters for the DPSO learning strategy and BC were also fixed to already published values (De-la Torre *et al.*, 2012a), and a sensitivity analysis was performed on the size of the LTM, picking the value that globally benefited the performance of the system (see Section 3.6). The weights and update discrimination parameters of the decision fusion were also fixed to the previously published values, and a sensitivity analysis

was conduced on the size of the LTM, finding that 30 frames is a good enough size (see Section 3.6).

3.5.1 Database for Face Re-Identification

Videos from the Carnegie Mellon University Face in Action (FIA) database were used in experiments (Goh *et al.*, 2005). It consists of 20-second videos captured at 30 frames per second, from 180 participants in a passport checking scenario. An array of 6 cameras was positioned at the face level to capture the scene, with a resolution of 640×480 pixels.. They are positioned at 0^o (frontal) and $\pm 72.6^o$ angle with respect to the individual. Three of the cameras were set at an 8-mm focal-length (zoomed), resulting in face areas around 300×300 pixels, and the other three at a 4-mm focal length (unzoomed) resulting in face areas around 100×100 pixels. Videos were captured in three sessions separated by a three-month interval for each subject. Zoomed cameras in all angles were used to retrieve enrollment/update trajectories, and the trajectories from unzoomed cameras were regrouped in a separated test set, and organized for the two experiments described below. Facial regions of interest (ROIs) were detected in videos using the well known Viola-Jones algorithm, using frontal, left and right profile according to the camera view (Viola and Jones, 2004).

Visual tracking was also applied on video sequences, initializing the Incremental Visual Tracker (IVT) (Ross *et al.*, 2008) with the first face detected, and tracking quality was stored for trajectory formation. All images were scaled to the highest possible resolution of the smallest face obtained after face detection (70x70 pixels). Features were extracted using Multi Scale Local Binary Patterns (MS-LBP) (Ojala *et al.*, 2002) with three block sizes (3×3, 5×5 and 9×9), along with pixel-intensity features. The resulting features were stacked in feature vectors, a PCA mapping was applied, and the 32 principal characteristics were selected.

Ten individuals of interest were randomly selected from the database, and one EoD was designed for each of them. Fig. 3.4 presents sample individuals in the two distinct scenarios considered in comparison: abrupt changes (pose) and gradual changes (age). However, the

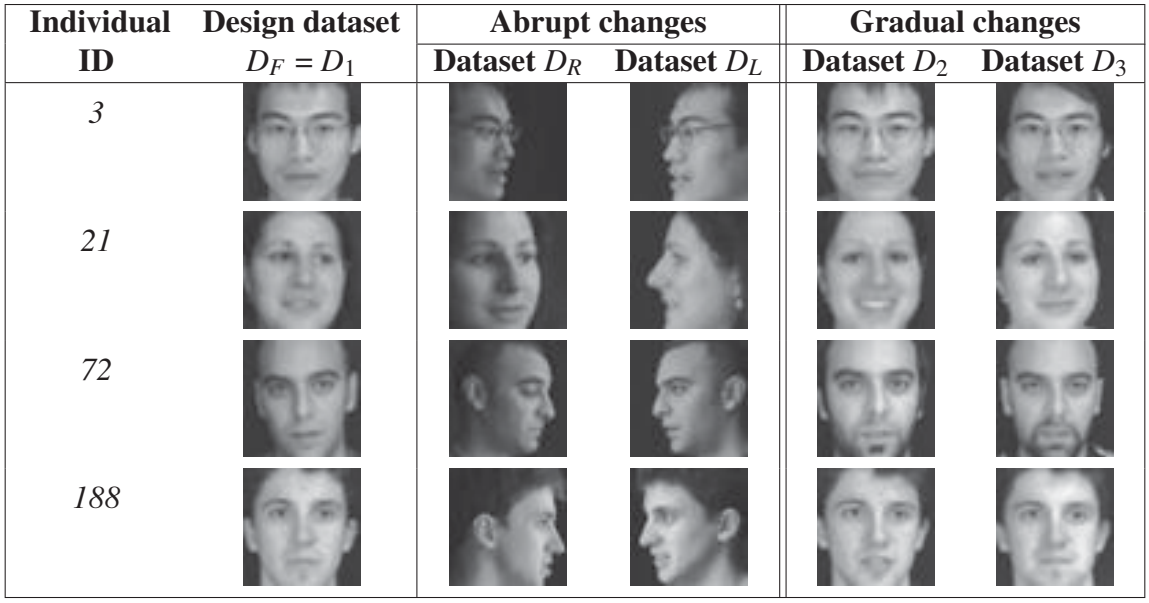


Figure 3.4 Samples of facial ROIs from 4 of the individuals of interest enrolled to the system. Faces were detected in video sequences from the FIA database using the Viola-Jones face detector trained with frontal faces for gradual changes, and frontal, right and left poses for abrupt changes

precise pose and age changes cannot be completely isolated, and in both scenarios some individuals also exhibit changes in makeup and expression. From the remaining individuals in the database, 88 were selected to build and maintain the universal model (UM), and the rest were considered as unknown individuals and appeared only on test. Note that in order to avoid a performance bias, the samples from individuals belonging to the UM do not appear in the test set, and similarly, samples from unknown individuals never appear on training stage.

During simulations, the amount of ROI samples retrieved from the trajectories for each individual of interest is shown on Table 3.3. The CM for individual k is comprised of 9 trajectories from non-target individuals in the cohort, and the number of ROI samples depends on the faces segmented for the corresponding trajectory. For example, in the scenario with abrupt changes, the reference ROI samples from trajectories in the CM of individual 2 are 1,474, 487 and 396 for the frontal, right and left datasets respectively. Similarly, the ROI samples from trajectories in the UM for the same scenario are 10,807, 1,713 and 3,205 from D_F, D_R and D_L respectively, extracted from 88 non-target trajectories in each block. Finally, the ROI samples in

the trajectories from unknown subjects ascend to 16,460, 2,678 and 5,081 for D_F, D_R and D_L respectively.

Table 3.3 Number of ROI samples in design and test trajectories for each individual of interest in the training and test datasets for both experiments. The system is designed with a single trajectory from D_F or D_1 (experiments 1 or 2 respectively), and updated twice with one trajectory from D_R (D_2) and D_L (D_3). The test set is composed of one trajectory from each pose for experiment 1, and from each capture session for experiment 2

ID	$\|T_k\|$ in experiment 1			$D_{tst-abrupt}$ (pose)			$\|T_k\|$ in experiment 2			$D_{tst-gradual}$ (age)		
(k)	$T_k \in D_F$	$T_k \in D_R$	$T_k \in D_L$	Front	Right	Left	$T_k \in D_1$	$T_k \in D_2$	$T_k \in D_3$	Front S1	Front S2	Front S3
2	149	54	80	114	30	73	149	208	184	114	109	119
3	170	62	54	194	31	23	170	149	123	194	151	165
21	92	61	38	75	31	36	92	138	184	75	79	101
58	202	57	11	176	6	64	202	254	202	176	215	172
72	223	38	41	144	22	58	223	268	246	144	184	151
99	82	53	40	57	72	43	82	146	0	57	115	0
121	126	63	48	68	47	27	126	122	113	68	57	46
188	148	46	72	118	66	52	148	183	233	118	172	192
190	190	65	60	132	18	59	190	217	148	132	92	88
213	241	42	32	110	39	48	241	210	234	110	83	85

3.5.2 Experimental Protocol

Prior to simulations, the design, update and test datasets were prepared for the two experiments, using trajectories extracted from FIA videos. The first experiment characterizes the system in a classification environment with abrupt changes (pose). Videos from the frontal, left and right cameras in the first capture session were used. The design set contains the enrollment trajectories from the frontal, zoomed camera (D_F). The trajectories from the right and left zoomed cameras were used to form the first and second update datasets, D_R and D_L respectively. The test set ($D_{tst-abrupt}$) is fixed, and contains trajectories from the frontal, right and left unzoomed views (poses).

The second experiment shows the behavior of the system in an environment with gradual changes, as propitiated by 3 months aging of the individuals. Here, the facial trajectories were extracted from videos recorded by the frontal cameras across the three capture sessions. The design set contains the enrollment trajectories from the first capture session, zoomed camera

(D_1). The trajectories from zoomed frontal view of the second and third sessions were used to form the first and second update datasets, D_2 and D_3 respectively. The test set ($D_{tst-gradual}$) is fixed, and contains trajectories from the frontal unzoomed view across the three capture sessions.

The CM and UM databases were maintained for each experiment, containing trajectories with similar characteristics than the design/update set. For instance, in the first experiment, the UM used for system design is composed of trajectories from the selected 88 individuals captured in the first session, with frontal, zoomed camera. In the second experiment, the CM at the first update stage is formed by trajectories of the non-target individuals of interest (enrolled to the system), captured in the second session, with frontal, zoomed camera. Non-target samples used for design and update are independently selected from the UM and CM for each training/validation set, using the proposed variant of One-Sided Selection.

The reference supervised incremental learning systems were first trained using trajectories from the design dataset, and new labeled reference samples were used for update. For these systems, it is assumed that the video sequences from the update datasets were manually labeled by an expert, and then used to update the system. In that sense, this scenario reflects the optimal case where the system is correctly updated anytime new reference samples become available, but also the most costly in terms of human effort. The reference supervised adaptive approaches include PFAM$_{i}nc$, Learn++(PFAM) and EoD$_{sup}$ (PFAM) $LTM_{KL,\lambda=\infty}$. These systems were updated with only the new labeled data. PFAM base classifiers were generated using the DPSO training algorithm, with an initial swarm of 60 particles, and a maximum of 5 particles within each of the 6 sub-swarms. The algorithm is set to run a maximum of 30 iterations, allowing 5 extra iterations to ensure convergence. Once the global best particle is found, its classifier as well as the 6 local best classifiers from each sub-swarm are added to the ensemble. The TCM-kNN was trained using $k = 1$, as published in (Li and Wechsler, 2005), and follows a batch learning scheme: on each update, past and new samples are learned from scratch. For instance, if a new trajectory becomes available at $t = 3$, the system is trained from scratch using $D_{batch} = D_1 \cup D_2 \cup D_3$.

Finally, the self-supervised system is first trained using the design set, and updated only when a trajectory T from unlabeled data blocks yields an accumulation that surpasses the update threshold, γ_k^{μ}. The approaches considered in this scenario include the EoD_{ss} (PFAM) with 6 different sizes of LTM: $\lambda = \{0, 25, 50, 75, 100, \infty\}$.

Learning is performed following a 2×5-fold cross - validation process for 10 independent trials. Positive samples from the incoming trajectory are randomly and evenly split in 5 folds of the same size. The folds are first distributed in three different design sets, including two folds for training (D^t), $1\frac{1}{2}$ fold to stop training epochs (D^e), and $1\frac{1}{2}$ fold for fitness evaluation (D^f). Once the classifiers are trained, D^e and D^f are combined, randomized and divided into two equally distributed subsets to produce a validation data to estimate a fusion function (D^c), and to select the operations point (D^s). Negative samples are chosen from the UM as well as the CM according to the proposed OSS+Rand selection strategy. Each fold is assigned to a different training/validation set for each replication of the experiment, and average performance measures are produced with five different assignments. At replication 5, the five folds are regenerated after a randomization of the sample order for each class.

3.5.3 Performance Analysis

The analysis of simulation results has been divided into three levels. First, *transaction-based analysis* shows the performance of the system based on classification decisions on each ROI. Then, a *subject-based analysis* allows a focus on specific individuals, which in turn allows for levels of performance depending on particular characteristics. Finally, a *trajectory-based analysis* shows the overall performance of the system after the decision fusion accumulates predictions for complete input trajectories (shown in Fig. 3.5).

Transaction-based performance analysis is used to assess the performance of the system for matching ROI samples to facial models. The true positive rate (tpr) and false positive rate (fpr) are estimated for different (fpr, tpr) operational points, and connected to draw a receiver operations characteristic (ROC) curve. When equal priors and costs are assumed, the closest

operations point to the upper-left corner corresponds to the optimal decision threshold. In applications with fpr constraint, the selection of the operations point is obtained from the graphical representation. The operations point is estimated on a validation subset used for operational predictions, providing a test (fpr,tpr) pair that reveals the generalization performance of the system at the selected point. The AUC (area under the curve) summarizes the performance depicted in a ROC graph, and the partial AUC (pAUC) focuses on a specific region of the curve, e.g. pAUC (5%) for an $fpr \leq 0.05$.

For different priors and costs of errors, the *Precision-Recall Operating Characteristic* (*PROC*) curve constitutes a graphical representation of detector performance where the impact of data imbalance is considered. The precision between positive predictions ($precision = TP/(TP + FP)$) is combined with the tpr (or *recall*) to draw a PROC curve. In general, the tpr is increased when the amount of positive (minority class) samples augments. On the contrary, the *precision* decreases with this amount. The scalar value of F_1-measure defined as $2 \cdot (precision \cdot tpr)/(precision + tpr)$ is used as a single performance indicator to combine recall and precision at a specific operations point.

It is well known that ensemble diversity has an impact in the performance of the ensemble, and the ambiguity is commonly used to measure diversity in ensembles (Zenobi and Cunningham, 2001). The ambiguity is defined by

$$\text{Ens. Ambiguity} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} amb(\mathbf{a}_n, d_m, d^*), \tag{3.24}$$

where M is the number of classifiers in the ensemble, N is the amount of test samples, and the ambiguity defined for an independent sample $\mathbf{a}_n$, given the decision d_m of the classifier c_m in the EoD, is given by

$$amb(\mathbf{a}_n, d_m, d^*) = \begin{cases} 0 & \text{if } d_m = d^* \\ 1 & \text{otherwise.} \end{cases} \tag{3.25}$$

The performance of FR systems may vary drastically from one person to the next, which is known as the "Doddington zoo" effect (Doddington *et al.*, 1998). In subject-based analysis, the error rates are assessed according to four different types of animals, rather than with the overall number of transactions. The resemblance of individuals performance to that of these animals can expose fundamental weaknesses in a biometric system, and allows the development of more robust systems. According to this characterization, the system tend to perform well in a *sheep*-like individual, irrespective of whether this individual belongs to the positive or negative class. *Goat*-like individuals belong to the positive class, but are difficult to identify, as they record consistently low classification scores against themselves. Goat-like individuals tend to determine the performance of the system through the disproportionate contribution to the false negative rate (fnr) of the system. A *wolf*-like individual belongs to the negative class, and is exceptionally successful at impersonating many different targets. Wolf-like individuals receive high scores when matched against others, and tend to elevate the false positive rate (fpr) of the system. Finally, a *lamb*-like individual is easy to impersonate, and thus seems usually susceptible to many different impostors. Lambs, on average, tend to produce high match scores when being matched against another user. For the last two cases, the match score distributions are significantly different from those of the general population.

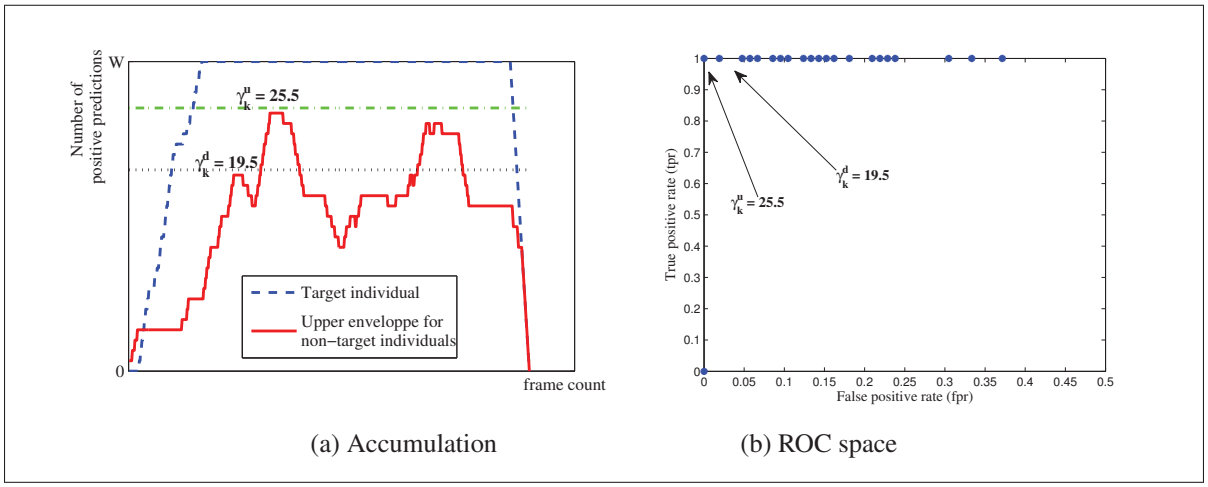
Table 3.4 Doddington's zoo thresholds for generalization performance at the operating point with $fpr = 1\%$, selected on validation data

Category	**Positive class**	**Negative class**
Sheep	$tpr \geq 55\%$ and not a lamb	$fpr \leq 1\%$
Lamb	At least 5% of non-target individuals are wolves	-
Goat	$tpr < 55\%$ and not a lamb	-
Wolf	-	$fpr > 1\%$

Typically, the likeliness of a user to one of the 4 aforementioned categories is defined at the score space. However, for binary classifiers, the confusion matrix can be used (Li and Wechsler, 2005). To establish a criterion, thresholds can be set at the fpr and fnr, and applied to

each EoD_k. Table 3.4 shows a criterion based on a system constraint of $fpr \leq 1\%$, considering a good fpr when it is just below 55%.

Trajectory-based performance analysis allows to assess performance over time of the entire system for person re-identification (see Fig. 3.2). All system functions are employed to process a video stream, including face detection, classification, tracking and spatio-temporal fusion. Indeed decisions taken by an operator occur on a time scale longer than a frame rate. Within the decision fusion system, positive predictions of each EoD_k are accumulated over a moving window of time for input ROI samples that correspond to a high quality facial track. Assume for instance a system that produces predictions at a maximum of 30fps. Each detected ROI is presented to all individual-specific EoDs of the system, which produces predictions (positive or negative) for each person enrolled to the system. Given a high quality face track, the number of positive predictions from an EoD should grow rapidly for the person of interest. Thus, the operator can more reliably detect the presence of a person of interest.



(a) Accumulation (b) ROC space

Figure 3.5 Trajectory-based analysis to evaluate the quality of a system for spatio-temporal FR in person re-identification

The adaptive system proposed in this chapter accumulates the positive predictions (responses of each EoD_k) over a window of W predictions. As shown in Fig. 3.5a, the quality of this system can be evaluated graphically by observing the evolution of positive predictions according to the

frame count (discrete time defined by the frame rate). In addition, once several individuals have appeared before of camera in a long video stream, and related trajectories have been processed, the quality of system decisions (i.e., the tpr, fpr, trr, frr) may be assessed over the range of decision threshold values, and represented in the ROC space (see Fig. 3.5b).

3.6 Results

Four strategies to select non-target training samples were compared in order to establish the most appropriate for the proposed system. The target samples in the datasets were maintained constant for all four strategies, and only the non-target samples were selected from CM and UM. A recently proposed CNN+Random method for the selection of non-target samples (De-la Torre *et al.*, 2014a) was used in comparison, as well as a distance based strategy, and the uniform random selection (Burghouts *et al.*, 2014). In general, the proposed OSS+Random selection of non-target samples permits to achieve a significantly higher level of performance than distance based and random selection alone in terms of F_1 at the selected operations point ($fpr = 1\%$), as shown in Figure 3.6. And although it presents similar performance than CNN+Random, its lower standard error makes OSS+Random selection a more desirable option. The OSS+Random strategy was used along all the simulations to select non-target samples for design and update of the EoDs.

Table 3.5 presents the average transaction-level performance obtained after design and update of the proposed and reference systems on ROI samples from trajectories stored in data blocks $D_F \rightarrow D_R \rightarrow D_L$, the scenario with abrupt changes. Measures used in comparison are the partial AUC for a $0 \leq fpr \leq 0.05$, pAUC (5%), as well as fpr, tpr and F_1 at a the operations point selected on the validation ROC curve for a desired $fpr = 1\%$. Performance for modular systems were measured for each individual-specific EoD, and average values over 10 individuals, 10 separated experiments (2×5 cross validation) are presented. The estimation of individual-specific ROC curves allows for a comparable performance measurement for the multi-class TCM-kNN. Note that the operations points and performance evaluation were computed after applying the rejection threshold provided by TCM-kNN, which is estimated during training.

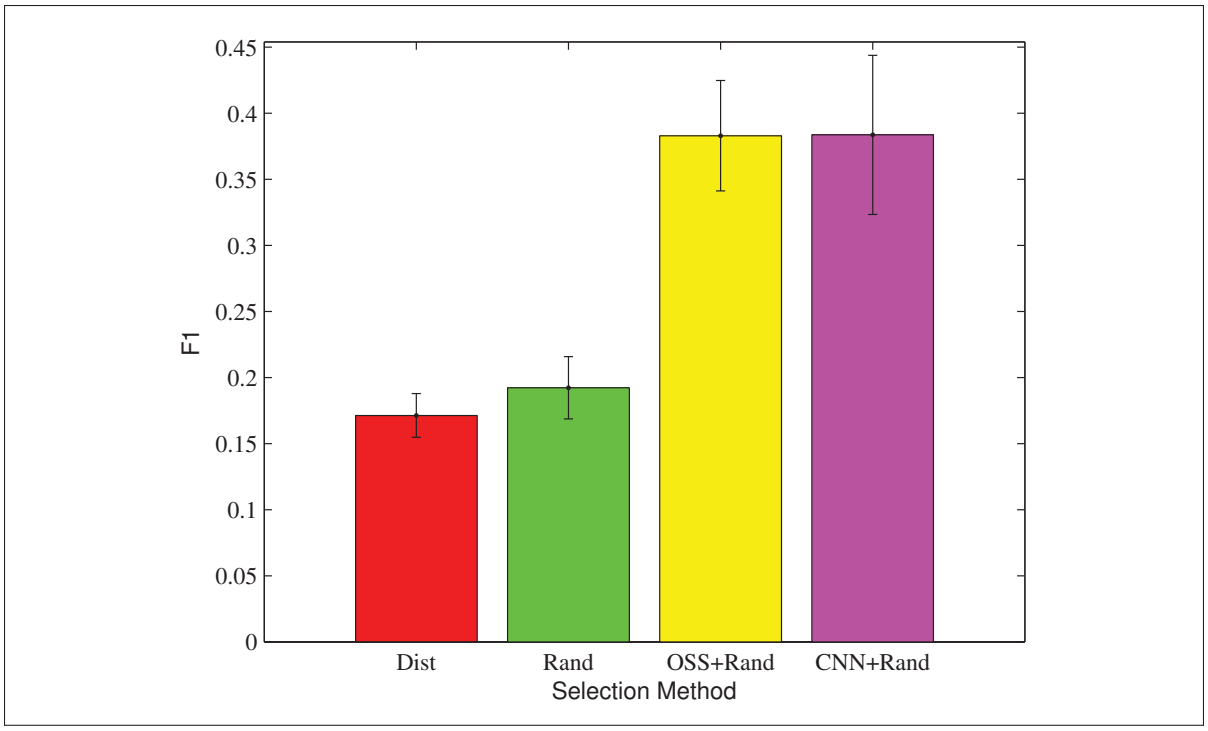


Figure 3.6 Comparison of different strategies to select training non-target samples from the CM and UM

Results on Table 3.5 show that TCM-kNN presents the poorest level of performance in terms of pAUC (5%) in the initial evaluation. After two updates, its level of performance is increased, but still remains significantly below all other approaches. Learn++ presents a pAUC (5%) significantly higher than TCM-kNN, however its performance consistently decreases after two updates. For PFAM$_{inc}$, the initial level of performance is higher than the last two approaches, and it also presents a performance decrease after updates, demonstrating the corruption of biometric models due to the abrupt changes in update trajectories. The EoD$_{sup}$ is the approach that presents the highest initial performance in terms of pAUC (5%), and it also presents a consistent increase that indicates its capacity to avoid the corruption of biometric models after each of the two updates. A similar trend is shown by all the approaches in terms of F_1 at the selected operating point. In addition, it can be observed that Learn++ and TCM-kNN allow to maintain a fpr close to the desired $fpr = 1\%$, at the expenses of a consistently low

Table 3.5 Average transaction-level performance over the 10 individuals of interest and for 10 independent experiments. Systems were designed-updated with $D_F \rightarrow D_R \rightarrow D_L$, and performance is shown after testing on $D_{test-abrupt}$, which involves ROIs from frontal, right and left views. In all cases, the operations point was selected using the ROC space on the validation dataset D^s at a $fpr = 1\%$, except for the partial AUC that comprises the operations points points for $0 \leq fpr \leq 0.05$

fpr (%) $\downarrow$	**tpr (%)** $\uparrow$	**precision (%)** $\uparrow$
• TCM-kNN		
1.40 ±0.44 $\rightarrow$ 0.67 ±0.23 $\rightarrow$ 0.71 ±0.23	5.59 ±1.69 $\rightarrow$ 5.44 ±1.65 $\rightarrow$ 7.72 ±2.33	0.50 ±0.15 $\rightarrow$ 1.10 ±0.36 $\rightarrow$ 1.33 ±0.42
• PFAM$_{inc}$		
2.72 ±0.35 $\rightarrow$ 5.31 ±0.68 $\rightarrow$ 4.44 ±0.74	50.81 ±1.65 $\rightarrow$ 55.13 ±3.12 $\rightarrow$ 55.08 ±3.28	31.92 ±2.32 $\rightarrow$ 26.48 ±2.77 $\rightarrow$ 26.41 ±2.52
• Learn++ (PFAM)		
1.00 ±0.09 $\rightarrow$ 1.26 ±0.32 $\rightarrow$ 0.50 ±0.08	15.99 ±1.95 $\rightarrow$ 11.32 ±2.12 $\rightarrow$ 3.68 ±0.98	12.07 ±1.06 $\rightarrow$ 13.23 ±1.70 $\rightarrow$ 7.26 ±1.54
• EoD$_{sup}$ (PFAM) LTM$_{KL,\lambda=\infty}$		
2.36 ±0.30 $\rightarrow$ 2.48 ±0.25 $\rightarrow$ 2.75 ±0.28	50.51 ±1.51 $\rightarrow$ 59.36 ±2.09 $\rightarrow$ 59.79 ±2.06	33.58 ±2.15 $\rightarrow$ 26.06 ±1.42 $\rightarrow$ 24.80 ±1.47
• EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda=\infty}$		
2.36 ±0.30 $\rightarrow$ 1.54 ±0.14 $\rightarrow$ 1.54 ±0.14	50.51 ±1.51 $\rightarrow$ 46.01 ±1.41 $\rightarrow$ 46.03 ±1.41	33.58 ±2.15 $\rightarrow$ 34.63 ±1.97 $\rightarrow$ 34.68 ±1.97

F_1 $\uparrow$	**pAUC (5%)** $\uparrow$
• TCM-kNN	
0.0091 ±0.0028 $\rightarrow$ 0.0177 ±0.0056 $\rightarrow$ 0.0223 ±0.0070	2.81 ±0.07 $\rightarrow$ 3.41 ±0.08 $\rightarrow$ 3.92 ±0.09
• PFAM$_{inc}$	
0.3281 ±0.0165 $\rightarrow$ 0.2344 ±0.0190 $\rightarrow$ 0.2641 ±0.0190	49.18 ±1.55 $\rightarrow$ 47.39 ±1.95 $\rightarrow$ 46.33 ±2.60
• Learn++ (PFAM)	
0.1204 ±0.0125 $\rightarrow$ 0.0731 ±0.0090 $\rightarrow$ 0.0274 ±0.0053	24.13 ±1.58 $\rightarrow$ 20.26 ±1.59 $\rightarrow$ 13.27 ±1.16
• EoD$_{sup}$ (PFAM) LTM$_{KL,\lambda=\infty}$	
0.3533 ±0.0163 $\rightarrow$ 0.3357 ±0.0143 $\rightarrow$ 0.3244 ±0.0147	53.16 ±1.38 $\rightarrow$ 58.09 ±1.50 $\rightarrow$ 64.31 ±1.58
• EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda=\infty}$	
0.3533 ±0.0163 $\rightarrow$ 0.3667 ±0.0150 $\rightarrow$ 0.3670 ±0.0150	53.16 ±1.38 $\rightarrow$ 56.18 ±1.27 $\rightarrow$ 56.18 ±1.27

$tpr < 20\%$. The opposite tendency is shown by PFAM$_{inc}$ and EoD$_{sup}$, which present around twice the desired fpr, but with a relatively high $tpr > 50\%$ that is maintained after update.

The proposed EoD$_{ss}$ maintains the capacity of avoiding knowledge corruption, inherited from EoD$_{sup}$, and also permits the reduction of manual labeling effort. Although it presents a slight reduction in pAUC (5%) with respect to EoD$_{sup}$, it is capable of enhanced performance in terms of F_1 at the selected operations point. This increase in performance can be explained by

the inherent increase in the diversity of the individual-specific EoDs after several self-updates with trajectories with different facial poses. And the fact that training and validation samples selected with the OSS+Random algorithm lie in the region of the feature space that causes most disagreement between base classifiers (Lu *et al.*, 2009). However, the performance of this system is also dependent on the amount of trajectories that the system correctly or incorrectly detected for self-update, number that highly variates from one individual to the other (see Table 3.1).

In a similar way, Table 3.6 presents the performance for design and update from trajectories in $D_1 \rightarrow D_2 \rightarrow D_3$ and test on $D_{tst-gradual}$, the scenario with gradual changes. Note that in this challenging scenario, all classifiers are initially trained using samples from frontal trajectories from the first capture session only, and are required to recognize samples from the three capture sessions. This scenario provides changes in expression, short term age (3 months) and distinct lookup like earrings, beard whiskers (see Fig. 3.4). Results in Table 3.6 show that TCM-kNN presents an increasing level of performance after each update, in terms of pAUC (5%), but this level still remains lower than other approaches, similar behavior presented in the scenario with abrupt changes. Learn++ produces an initial level of performance that is similar to the level shown in the scenario with abrupt changes, and the pAUC (5%) after two updates reveals that the models were also affected by knowledge corruption. The level of performance presented by the EoD$_{sup}$ in terms of pAUC (5%) is also superior to other approaches. And the self-update strategy of the EoD$_{ss}$ allows to increase the level of performance after update on D_2, but this level suffers a slight decrease after update on D_3. This reduction is also related to the amount of trajectories that the system correctly or incorrectly employed for self-update, which affect differently to each individual-specific EoD (see Tables 3.9 and 3.2). In fact, the amount of non-target trajectories wrongly used for the first self-update was superior in the scenario with gradual changes with respect to the scenario with abrupt changes (261 vs. 181 wrong updates). And most of the wrong updates in the scenario with abrupt changes were performed by the EoD$_{ss}$ from a single individual (111 wrong updates from individual 58).

Table 3.6 Average transaction-level performance over the 10 individuals of interest and for 10 independent experiments. Systems were designed-updated with $D_1 \rightarrow D_2 \rightarrow D_3$, and performance is shown after testing on $D_{test-gradual}$, which involves frontal ROIs from the first, second and third capture sessions. In all cases, the operations point was selected using the ROC space on the validation dataset D^s at a $fpr = 1\%$, except for the partial AUC that comprises the operations points for $0 \leq fpr \leq 0.05$

fpr (%) $\downarrow$	**tpr (%)** $\uparrow$	**precision (%)** $\uparrow$
• TCM-kNN		
2.18 ±0.67 $\rightarrow$ 1.99 ±0.62 $\rightarrow$ 1.97 ±0.61	7.46 ±2.25 $\rightarrow$ 9.46 ±2.86 $\rightarrow$ 9.90 ±2.98	0.39 ±0.12 $\rightarrow$ 0.55 ±0.17 $\rightarrow$ 0.57 ±0.18
• PFAM$_{inc}$		
2.43 ±0.22 $\rightarrow$ 3.01 ±0.31 $\rightarrow$ 3.30 ±0.36	64.25 ±2.33 $\rightarrow$ 74.54 ±2.50 $\rightarrow$ 73.01 ±2.36	30.11 ±1.67 $\rightarrow$ 35.93 ±2.42 $\rightarrow$ 34.57 ±2.34
• Learn++ (PFAM)		
1.01 ±0.08 $\rightarrow$ 1.67 ±0.23 $\rightarrow$ 1.97 ±0.21	16.03 ±2.36 $\rightarrow$ 19.94 ±2.39 $\rightarrow$ 19.17 ±2.69	12.16 ±1.17 $\rightarrow$ 12.69 ±1.35 $\rightarrow$ 8.40 ±0.81
• EoD$_{sup}$ (PFAM) LTM$_{KL,\lambda=\infty}$		
2.28 ±0.19 $\rightarrow$ 2.18 ±0.20 $\rightarrow$ 2.43 ±0.39	63.92 ±2.41 $\rightarrow$ 61.93 ±2.58 $\rightarrow$ 60.09 ±2.62	29.83 ±1.46 $\rightarrow$ 30.00 ±1.70 $\rightarrow$ 30.92 ±1.80
• EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda=\infty}$		
2.28 ±0.19 $\rightarrow$ 5.44 ±0.61 $\rightarrow$ 4.95 ±0.64	63.92 ±2.41 $\rightarrow$ 63.63 ±2.78 $\rightarrow$ 55.19 ±2.93	29.83 ±1.46 $\rightarrow$ 21.05 ±1.82 $\rightarrow$ 24.46 ±2.32

F_1 $\uparrow$	**pAUC (5%)** $\uparrow$
• TCM-kNN	
0.0073 ±0.0022 $\rightarrow$ 0.0103 ±0.0032 $\rightarrow$ 0.0108 ±0.0034	3.94 ±0.07 $\rightarrow$ 4.45 ±0.05 $\rightarrow$ 4.59 ±0.04
• PFAM$_{inc}$	
0.3662 ±0.0125 $\rightarrow$ 0.4051 ±0.0172 $\rightarrow$ 0.4029 ±0.0187	66.55 ±1.85 $\rightarrow$ 73.58 ±1.85 $\rightarrow$ 73.26 ±1.78
• Learn++ (PFAM)	
0.1170 ±0.0128 $\rightarrow$ 0.1312 ±0.0131 $\rightarrow$ 0.1040 ±0.0118	24.90 ±1.93 $\rightarrow$ 25.70 ±2.14 $\rightarrow$ 19.76 ±1.98
• EoD$_{sup}$ (PFAM) LTM$_{KL,\lambda=\infty}$	
0.3664 ±0.0121 $\rightarrow$ 0.3652 ±0.0146 $\rightarrow$ 0.3620 ±0.0153	70.83 ±1.56 $\rightarrow$ 81.78 ±0.97 $\rightarrow$ 83.24 ±0.97
• EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda=\infty}$	
0.3664 ±0.0121 $\rightarrow$ 0.2599 ±0.0144 $\rightarrow$ 0.2710 ±0.0184	70.83 ±1.56 $\rightarrow$ 77.15 ±1.35 $\rightarrow$ 75.27 ±1.53

Finally, the experiment was repeated with 10 different lists of randomly selected individuals to design 10 different modular systems. The system was characterized in both test scenarios with gradual and abrupt changes. Since user-specific analysis shows that the system behaves differently from individual to individual (see Section 3.6.1), this variation in the experiment allows to discard the bias induced by the initially selected group of individuals of interest. The average performance for the system over all the lists of randomly selected individuals of interest is shown in Table 3.7. Looking at the results for scenario with abrupt changes (first

Table 3.7 Average transaction-level performance over the 10 different systems designed for 10 randomly selected individuals of interest each. In the first case (top row), the systems were designed-updated with $D_F \rightarrow D_R \rightarrow D_L$, and performance is shown after testing on $D_{test-abrupt}$, which involves ROIs from frontal, right and left views. In the second case (bottom row), the systems were designed-updated with $D_1 \rightarrow D_2 \rightarrow D_3$, and performance is shown after testing on $D_{test-gradual}$, which involves frontal ROIs from the first, second and third capture sessions. In all cases, the operations point was selected using the ROC space on the validation dataset D^s at a $fpr = 1\%$, except for the partial AUC that comprises the operations points for $0 \leq fpr \leq 0.05$

fpr (%) ↓	**tpr (%)** ↑	**precision (%)** ↑	**F_1** ↑	**pAUC (5%)** ↑
Abrupt Changes ($D_F \rightarrow D_R \rightarrow D_L$)				
• EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda=\infty}$				
2.38 ±0.64 → 2.40 ±0.65 → 1.94 ±0.56	8.92 ±2.72 → 8.91 ±2.70 → 8.41 ±2.62	6.06 ±2.42 → 6.05 ±2.39 → 7.22 ±2.73	0.0587 ±0.0176 → 0.0586 ±0.0177 → 0.0617 ±0.0186	12.90 ±3.16 → 12.10 ±2.96 → 20.66 ±5.76
Gradual Changes ($D_F \rightarrow D_R \rightarrow D_L$)				
• EoD$_{ss}$ (PFAM) LTM$_{KL,\lambda=\infty}$				
17.72 ±5.94 → 12.91 ±6.05 → 8.28 ±4.71	22.21 ±10.24 → 9.03 ±7.07 → 11.94 ±7.25	1.14 ±0.59 → 2.29 ±1.43 → 2.66 ±1.40	0.0214 ±0.0111 → 0.0240 ±0.0137 → 0.0316 ±0.0144	13.57 ±3.65 → 49.94 ±11.08 → 50.96 ±10.81

row of the Table 3.7), it can be seen that the pAUC (5%) performance of the system slightly drops after one self-update step, but is increased after two self-updates, reaching a level that is higher than the initial performance. This confirms the tendency observed in Table 3.5 for the different performance measures. Similarly, when the trajectories used in self-update present gradual changes (second row of the Table 3.7), the self-update system shows an increase in the level of performance in terms of pAUC (5%). This is also consistent with the tendency found in the results from Table 3.6, confirming the behavior of the system for the different individuals in the FIA database.

Different conclusions can be made from the comparison of the performances achieved in the two scenarios (Tables 3.5, 3.6 and 3.7). Regarding the supervised approaches, the pAUC (5%) of TCM-kNN increases after each update, although it is in general lower than all other approaches, even though all reference samples are stored in the facial model. This phenomenon is related to the difficulty faced by multi-class classifiers in finding multiple boundaries, as opposite to 2-class classifiers that only divide the feature space in two regions. On the contrary, the pAUC (5%) of the Learn++ approach decreases after 2 updates in both scenarios, although the same data and training strategy was used to design its base classifiers. PFAM$_{inc}$ is successful adapting to gradual changes: when update trajectories contain ROIs that are very similar to

those originally used for design. However, when update trajectories significantly differ from training trajectories (abrupt changes), PFAM$_{inc}$ suffers of knowledge corruption. As expected, the EoD$_{sup}$ allows to alleviate the knowledge corruption presented by PFAM$_{inc}$ by using the *learn-and-combine* strategy, and in general it achieves the highest pAUC (5%). And the EoD$_{ss}$ allows to improve the performance of facial models in both scenarios in terms of pAUC (5%). However, in the scenario with gradual changes it present difficulties at selecting an operations point that generalizes the performance, which is reflected in a decrease in F_1. These difficulties are originated because training and validation samples for design and update in the scenario with gradual changes are very similar, which biases the operations point selection and causes the overtraining of the ensemble.

Besides, it is well known that the diversity of opinions in an ensemble is correlated with the final accuracy of the ensemble (Kuncheva, 2004), and the ambiguity in the scenario with gradual changes decreases after each update (see Fig. 3.7). Thus, when the changes in the environment are gradual, it is preferable to weaken the learning strategy of the EoD$_{ss}$ by reducing the amount of base classifiers learned on each adaptation. And a scenario with abrupt changes is better addressed by the proposed EoD$_{ss}$.

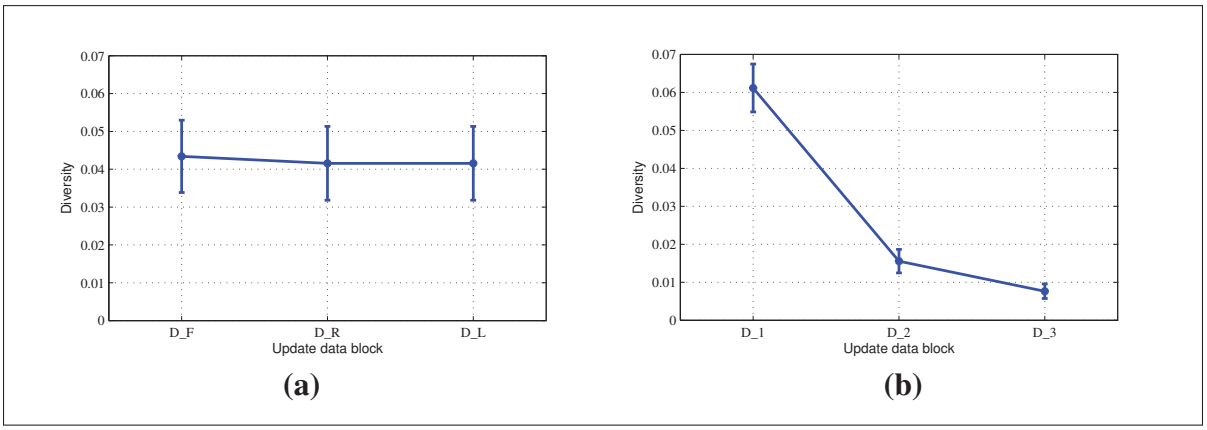


Figure 3.7 Evolution of the average ensemble ambiguity of the EoD$_{ss}$ after each update in the scenarios with abrupt changes (a), and gradual changes (b)

3.6.1 Subject-Based Analysis

In order to proceed with the subject-based analysis, four individuals were selected by their characterization under the Doddington zoo terminology, after their initial design/test cycle, and test on $D_{test-abrupt}$. According the criteria established on Table 3.4, the EoD$_{ss}$ (3) corresponds to a sheep-like individual. For this individual, 5 of the 146 unknown subjects in the test set are recognized as targets more than 1% of the time (wolves), and present an average $fpr > 1\%$. Similarly, the EoD$_{ss}$ (21) and EoD$_{ss}$ (188) correspond to lamb-like individuals, with 37 and 51 wolves respectively, and a $fpr > 1\%$ in both cases. The EoD$_{ss}$ (72) corresponds to a goat-like individual, with 28 wolves and an $fpr > 1\%$.

Table 3.8 presents the average subject-based performance for the 4 individuals of interest in the scenario with abrupt changes. Even though the EoD$_{ss}$ for individual 3 presented the initial highest performance in terms of pAUC (5%) and F_1 when compared to the EoD$_{ss}$ from other individuals, the system was never updated on ROIs from the update sets (D_R and D_L). The low level of fpr on test indicates that the EoD$_{ss}$ rejects very well the non-target ROIs, and the relatively high tpr indicates that a high amount of ROIs were correctly recognized as target. However, the self-update mechanism is never activated, and the reason can be easily explained by observing the accumulated responses from the EoD$_{ss}$ (3) shown in Fig. 3.8.

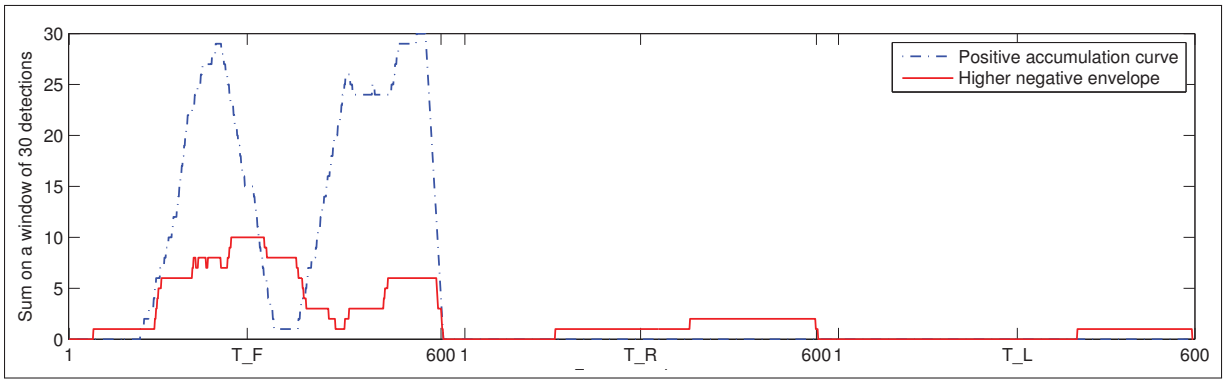


Figure 3.8 Example of accumulated responses of the EoD$_{ss}$ (3) after design on D_F, and test on frontal, right and left trajectories from $D_{test-abrupt}$, which includes pose changes

The curves in Fig. 3.8 show that a high amount of ROIs from the trajectory from the frontal view can be correctly detected by the EoD$_{ss}$ (3), but none of the target ROIs from right and left poses. Recalling the number of ROIs retrieved from individual 3 in $D_{test-abrupt}$ (see Table 3.3), around 78% of them correspond to the frontal trajectory, and constitute bias towards frontal facial captures. These observations evidence a weakness of the transaction-based analysis applied to spatio-temporal systems, and encourage the use of trajectory-based evaluation.

Table 3.8 Average performance of the system for 4 individuals of interest over 10 independent experiments, after design/update on $D_F \rightarrow D_R \rightarrow D_L$

EoD$_{ss}$ (3) (*sheep*-like)	**EoD$_{ss}$ (21)** (*lamb*-like)	**EoD$_{ss}$ (188)** (*lamb*-like)	**EoD$_{ss}$ (72)** (*goat*-like)
fpr (%) $\downarrow$			
0.38 ±0.08 → 0.37 ±0.07 → 0.37 ±0.07	4.97 ±0.97 → 4.24 ±2.31 → 4.21 ±2.31	1.61 ±0.83 → 5.53 ±2.60 → 5.50 ±2.60	1.94 ±0.59 → 1.78 ±0.37 → 1.78 ±0.37
tpr (%) $\uparrow$			
56.13 ±3.79 → 56.77 ±3.77 → 56.77 ±3.77	58.45 ±2.68 → 35.99 ±7.02 → 35.77 ±7.03	49.93 ±3.63 → 49.79 ±3.94 → 49.92 ±3.87	55.80 ±1.86 → 51.74 ±2.91 → 51.74 ±2.94
Precision (%) $\uparrow$			
70.05 ±2.99 → 70.18 ±2.83 → 70.18 ±2.83	12.11 ±2.80 → 13.89 ±4.12 → 13.88 ±4.23	37.28 ±6.12 → 20.23 ±3.47 → 20.46 ±3.48	36.03 ±5.61 → 34.53 ±5.99 → 34.47 ±6.01
F$_1$ $\uparrow$			
0.611 ±0.025 → 0.617 ±0.025 → 0.617 ±0.025	0.185 ±0.028 → 0.179 ±0.047 → 0.179 ±0.048	0.383 ±0.042 → 0.272 ±0.036 → 0.274 ±0.036	0.410 ±0.038 → 0.389 ±0.041 → 0.390 ±0.042
***p*AUC (5%)** $\uparrow$			
71.31 ±1.20 → 71.31 ±1.20 → 71.31 ±1.20	48.28 ±3.34 → 47.64 ±2.60 → 47.64 ±2.60	51.33 ±2.10 → 51.87 ±2.53 → 51.87 ±2.53	59.04 ±1.05 → 58.17 ±1.13 → 58.17 ±1.13

The two analyzed lamb-like individuals are specially interesting given that each of them presents a different affectation in their performance. Individual 21 shows an initial high tpr, which is negatively affected as the system performs self-update, whereas the fpr remains high but almost constant. This means that the EoD$_{ss}$ (21) maintains a robust rejection against non-target samples, but is weak maintaining the level of target ROIs. On the other hand, individual 188 presents a relatively low fpr, and it increases as the system performs the self-update, maintaining a tpr almost constant. Thus, the EoD$_{ss}$ (188) maintains its robustness in detecting target trajectories, but is weak maintaining the rejection capacity against non-target. From this observations, it can be concluded that false updates can affect differently to distinct lamb-like individuals, remarking the need for an individual independent characterization of the system.

Figure 3.9 shows the accumulated decisions of the EoD$_{ss}$ for individuals 21 and 188. The curves produced by the EoD$_{ss}$ (21) for the trajectories in $D_{tst-abrupt}$ show that this system

is capable of a correct discrimination between target and non-target frontal trajectories, but present difficulties recognizing the face of the individual captured with different poses, specially from the left view. As a consequence, the EoD_{ss} (21) is often correctly updated with right pose trajectories, but the incorrect updates are also common. The curves for EoD_{ss} (188) show that the system is capable of correctly detect target trajectories from the frontal and right views, but also wrongly detect non-target trajectories. And it fails to detect target trajectories from the left view, but wrongly detects non-target trajectories from the left view.

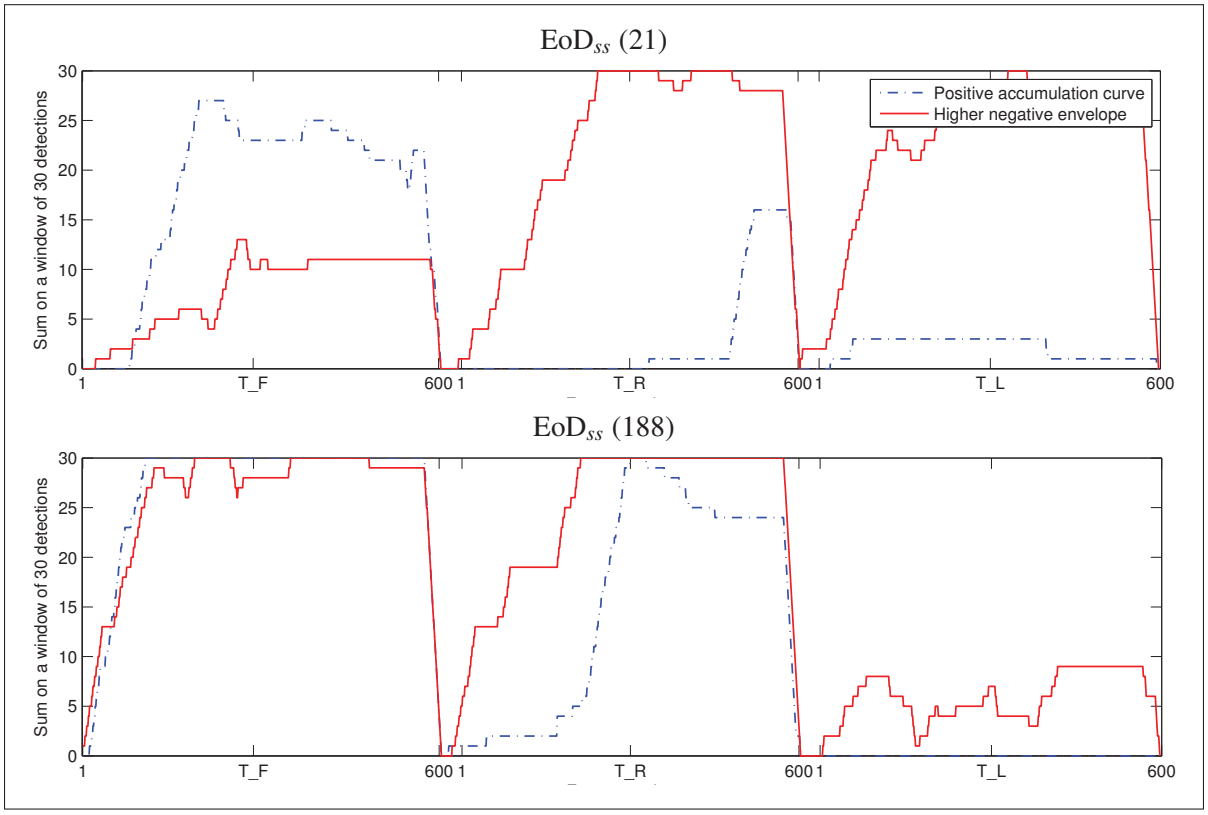


Figure 3.9 Example of accumulated responses of the EoD_{ss} for the lamb-like individuals, after design on D_F, and test on frontal, right and left trajectories from $D_{test-abrupt}$, which includes pose changes

Finally, the goat-like individual 72 maintains over time a relatively constant fpr that is not significantly affected by its wolf-like individuals (samples from wolf-like individuals represent less than 1% of the fpr, $fpr < 1\%$). It also shows a low tpr that avoids any correct self-updates

by the system (see Table 3.1). However, the EoD$_{ss}$ (72) is capable to maintain the low level of fpr even though it presented several wrong updates.

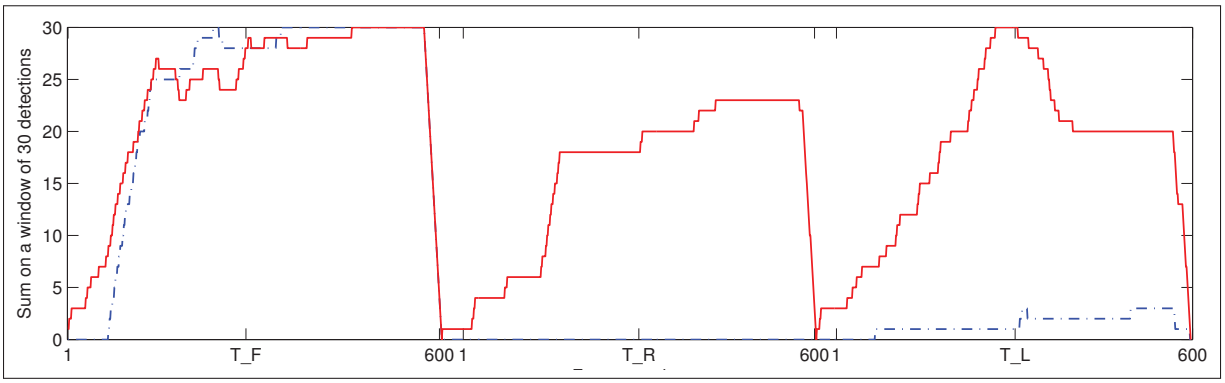


Figure 3.10 Example of accumulated responses of the EoD$_{ss}$ for the goat-like individual 72

Figure 3.10 presents the accumulated responses for the trajectories in $D_{test-abrupt}$, showing the difficulty of the system to differentiate between target and non-target trajectories. The EoD$_{ss}$ (72) successfully detects frontal target trajectories but presents difficulties detecting target trajectories from the left and right view, and it wrongly detects several non-target trajectories from all views.

Table 3.9 presents the individual-specific average performance of ensembles for the semi-supervised scenario obtained after self-update using ROI samples from trajectories stored in D_1, D_2 and D_3. According to these results, the EoD$_{ss}$ (3) allows to maintain the initial level of performance after two updates, although it is lower in terms of F_1 score. This trend is similar as the observed in the scenario with abrupt changes, but in this case is product of several correct and incorrect self-updates. The EoD$_{ss}$ (21) maintained the level of fpr after two self-updates, but the tpr decreased significantly, also a similar trend as shown in the scenario with abrupt changes. In the same way, for the EoD$_{ss}$ (188) the fpr was augmented and the tpr reduced as product of multiple false updates. And differently from the scenario with gradual changes, the EoD$_{ss}$ (72) presents an increasing fpr and decreasing tpr, produced by multiple wrong self-updates. According to these observations, the biometric models for individuals 3

and 21 were slightly affected by wrong self updates. For individuals 188, the biometric model was corrupted by wrong self-updates, and follows the same trend as the scenario with abrupt changes. This trend is related to wrongly detect an increasing number of non-target individuals which eventually will make it useless. The biometric model of individual 72 was also damaged by wrong self-updates, shows a similar decreasing performance trend, and is reinforced by its initial low performance.

Table 3.9 Average performance of the system for 4 individuals of interest over 10 independent experiments, after design/update on $D_1 \rightarrow D_2 \rightarrow D_3$

EoD$_{ss}$ (3)	**EoD$_{ss}$ (21)**	**EoD$_{ss}$ (188)**	**EoD$_{ss}$ (72)**
fpr (%) $\downarrow$			
0.69 ±0.21 → 3.11 ±2.30 → 1.41 ±0.97	3.12 ±0.53 → 4.75 ±1.56 → 3.37 ±1.21	4.77 ±1.11 → 12.05 ±2.76 → 10.84 ±3.78	2.12 ±0.39 → 6.83 ±2.00 → 8.08 ±2.43
tpr (%) $\uparrow$			
39.24 ±5.95 → 38.35 ±4.04 → 37.82 ±7.05	73.57 ±3.86 → 76.27 ±4.51 → 65.06 ±5.99	95.27 ±2.00 → 98.22 ±0.62 → 83.24 ±8.00	34.03 ±2.52 → 37.29 ±7.51 → 28.81 ±4.50
Precision (%) $\uparrow$			
55.48 ±5.21 → 52.41 ±9.05 → 60.51 ±8.06	20.76 ±3.45 → 17.97 ±3.25 → 29.00 ±6.94	30.38 ±4.14 → 18.18 ±4.24 → 26.00 ±7.61	23.61 ±2.63 → 12.66 ±3.05 → 8.72 ±2.67
F$_1$ $\uparrow$			
0.424 ±0.039 → 0.368 ±0.043 → 0.422 ±0.063	0.302 ±0.031 → 0.270 ±0.035 → 0.342 ±0.061	0.443 ±0.045 → 0.288 ±0.056 → 0.347 ±0.081	0.263 ±0.013 → 0.170 ±0.036 → 0.124 ±0.032
***p*AUC (5%)** $\uparrow$			
71.89 ±3.84 → 76.23 ±3.97 → 74.30 ±4.82	70.59 ±1.98 → 77.51 ±2.00 → 79.23 ±2.21	90.13 ±1.96 → 93.13 ±1.32 → 89.10 ±3.01	38.87 ±1.78 → 53.72 ±4.82 → 45.86 ±5.19

In summary, although the initial performance for the sheep-like individual 3 was very similar in terms of *p*AUC (5%) with respect to the gradual changes scenario, it was increased after two self-updates. The lamb-like individuals 188 presented a slight degradation in performance produced by several wrong self-updates, similar trend with respect to the scenario with abrupt changes. But the performance for individual 21 was increased in terms of pAUC (5%), indicating that the wrong and correct self-updates allowed to increase the diversity of the EoD$_{ss}$ (21) without affecting its accuracy. Finally, the level of performance for the EoD$_{ss}$ (72) shows an increase in terms of *p*AUC (5%), but a constant decrease in terms of F_1 score. This tendency differs from the scenario with abrupt changes, and can be explained by the multiple wrong self-updates compared with the few correct self-updates (see Table 3.2).

3.6.2 LTM management

Figure 3.11 presents the average performance in terms of F_1 score for different sizes of LTM (λ_k values) used by EoD$_{ss}$. The graph shows the performance for the whole system, as well as the average for the 2 lamb-like individuals analyzed before. The average for the whole system shows a constant increase as λ_k increases, supporting that more data in the LTM allows for a more accurate system. However, as shown by individual-specific graphs, this affirmation is different for each individual. The EoD$_{ss}$ (21) shows the highest performance when $\lambda_{21} = \infty$, but the performance does not increases in the same manner as the value of λ_{21}. An increase in the performance for $\lambda = 25, 50$, and a decrease for $\lambda = 75, 100$ shows that the LTM management strategy allows to filter out some non-useful samples that negatively affect the performance. On the other hand, the EoD$_{ss}$ (188) presents its peak performance when $\lambda_{188} = 100$, confirming that some non-useful samples were discarded.

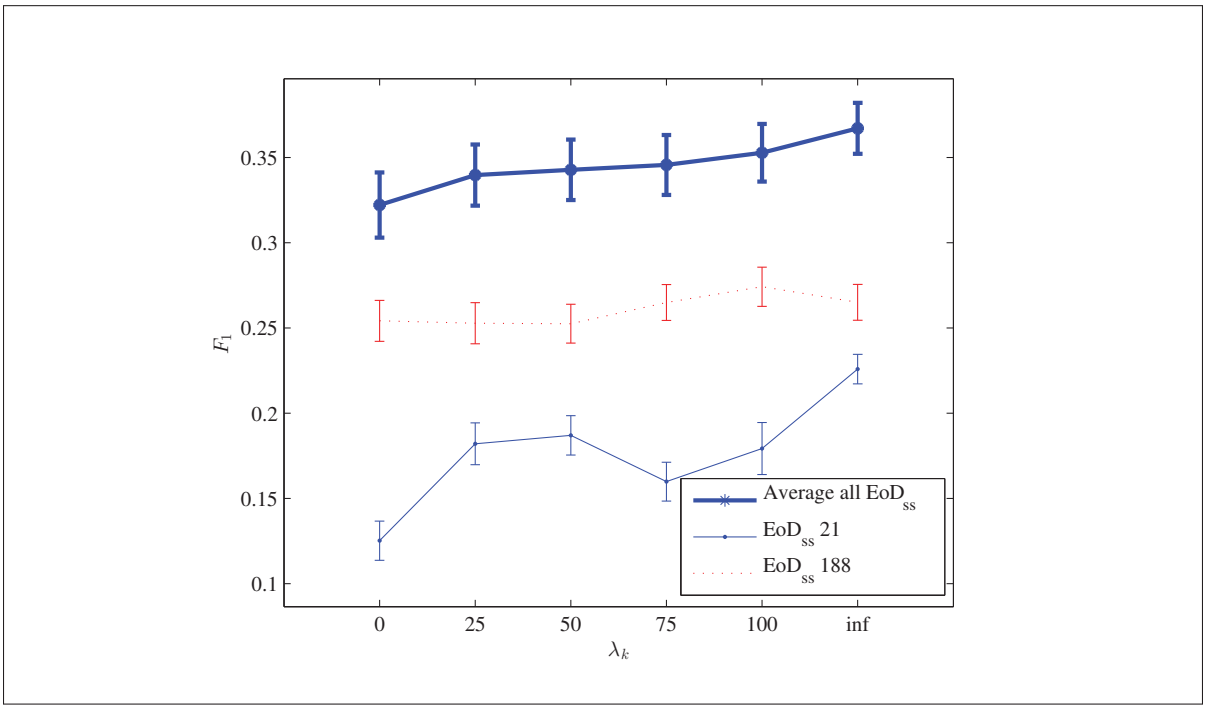


Figure 3.11 Performance in terms of F_1 at the operations point with $fpr = 1\%$. Average for all individuals and the EoD$_{ss}$ for the lamb-like individuals with ID 21 and 188

Preserving validation samples from wolf-like individuals is crucial to reduce or maintain a low fpr as the system is self-updated. Specially for lamb-like individuals, for which the fpr performance is affected by wolf-like individuals, and the proposed LTM management scheme allows to rank and select these samples to be retained over time. Fig. 3.12 shows the percentage of samples from wolf-like individuals for a LTM with different sizes ($\lambda_k = [1...100]$). The three different ranking measures that were compared are the KL divergence, average margin sampling (AMS) and vote entropy. Results shown in Fig. 3.12 reveal that both, the KL divergence and vote entropy, enable the system to select the highest amount of samples from wolf-like individuals. And is the KL divergence the measure that consistently shows the highest percentage for small LTM sizes ($\lambda_k \leq 20$).

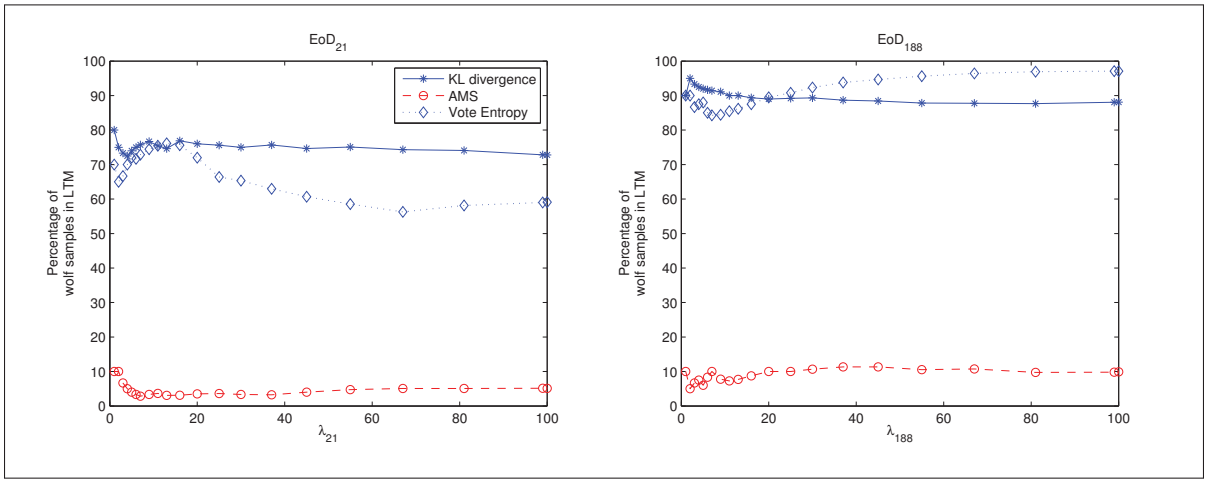


Figure 3.12 Percentage of samples from wolf-like individuals for the EoD_{ss} for the lamb-like individuals with ID 21 and 188

3.6.3 Trajectory-Based Analysis

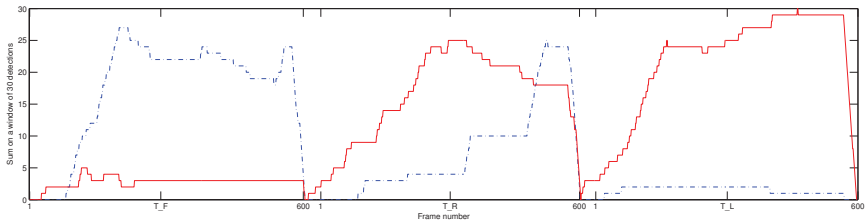
Different decision fusion methods were compared in order to establish the cases where the proposed scheme is strong. Fig. 3.13 presents an example of the sequential output produced by the system after the input of three different trajectories. The proposed method that accumulates decisions, either over a window of frames or detections, is presented at the top of the Fig. 3.13,

followed by the raw scores that are commonly used in literature (Despiegel *et al.*, 2012). Then the accumulated scores over a fixed size window of frames serves as a reference point for comparison, as well as the sum of the maximum scores provided by the EoD_{ss} (Ekenel *et al.*, 2010).

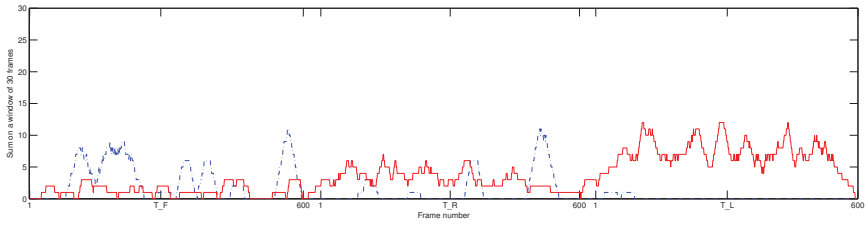
Even though the graphs in Fig. 3.13 correspond to a single experiment (replication 5 for individual 21), they allow to observe the behavior of the different decision fusion techniques. Now, the response of the system to the easiest trajectory for the EoD_{ss} can be analyzed, i.e. T_F with frontal views as seen before. The accumulated decisions for T_F provide a wide separation between accumulations for the target and non-target trajectories. A similar trend is shown by the accumulation of decisions over a window of frames, with less separation, followed by the accumulation of scores with an even smaller separation between target and non-target trajectories. For the raw scores fusion, the separation cannot be easily established, and for this case, the sum of maximum scores does not provides any discrimination given that non-target trajectories produce higher accumulations than the sum of target scores.

Figure 3.14 presents the average trajectory-based ROCs for 10 individuals, 10 replications, in the two scenarios, using the different methods for decision fusion. These results confirm that the proposed fusion strategy allows for higher discrimination between trajectories.

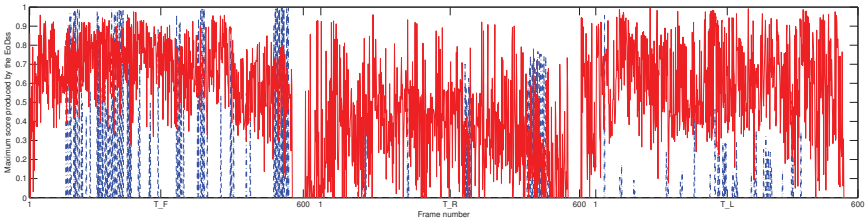
A test on the sensitivity of the system to different window sizes was performed for the both analyzed scenarios, comparing distinct decision fusion methods. Fig. 3.15 presents the average pAUC(5%) for the curves after applying different thresholds to the trajectories from $D_{tst-abrupt}$ and $D_{tst-gradual}$. Reference decision fusion methods compared include using the maximum score on each trajectory (Despiegel *et al.*, 2012), the accumulation of maximum scores over a time window, and the sum of all scores over consecutive frames. As expected, the performance of the single-score decision scheme is not affected by the size of the window. The accumulation of scores scores over a window of frames improves the performance of the single-score approach, but the sum over all the frames in the trajectory is slightly better, as shown in Fig. 3.15. The proposed accumulation on decisions for each facial detection is superior to all other



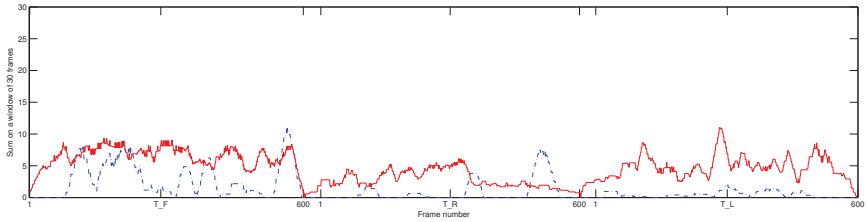
(a) Accumulated decisions on a fixed size window of detections



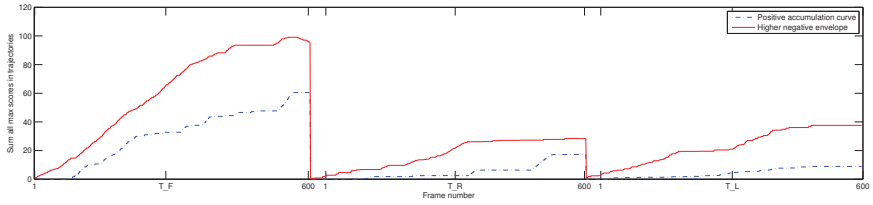
(b) Accumulated decisions on a fixed size window of frames



(c) Raw scores on a fixed size window of frames



(d) Accumulated scores on a fixed size window of frames



(e) Sum of the maximum scores for all past frames

Figure 3.13 Examples of evolution curves for different decision fusion methods

approaches in terms of pAUC(5%), as can be shown in the Fig. 3.15. And the accumulation over a window of frames allows co-jointly evaluate the performance of the whole system at once, including not only the tracker and classifier, but also the face segmentation algorithm.

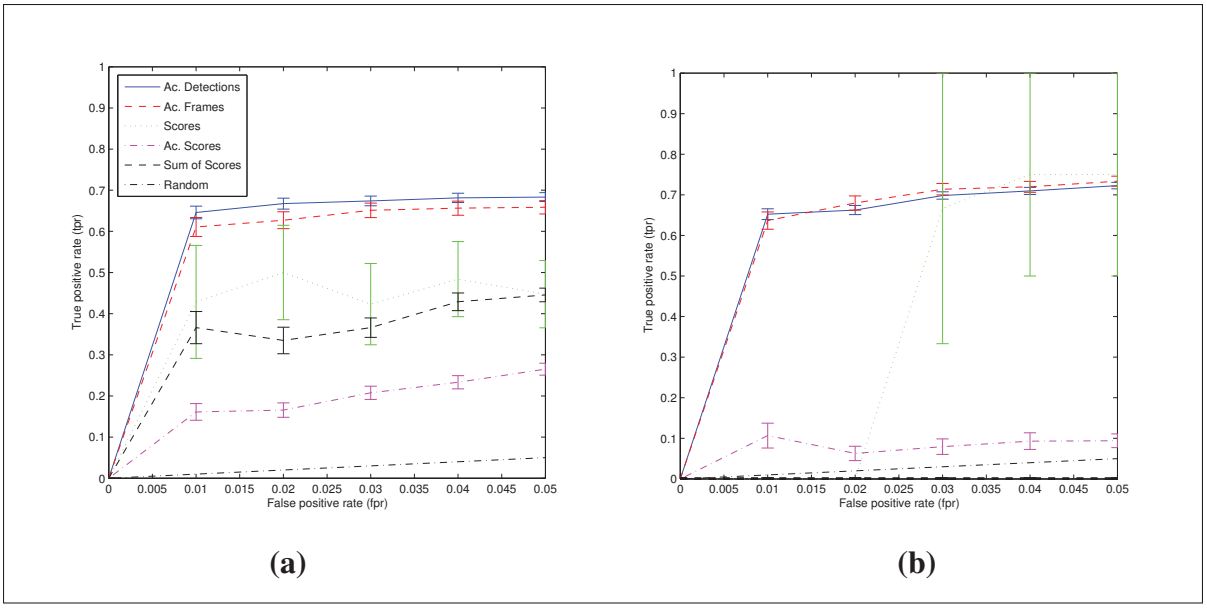


Figure 3.14 Average global ROCs for the system after update in the scenarios with abrupt changes (a) and gradual changes (b)

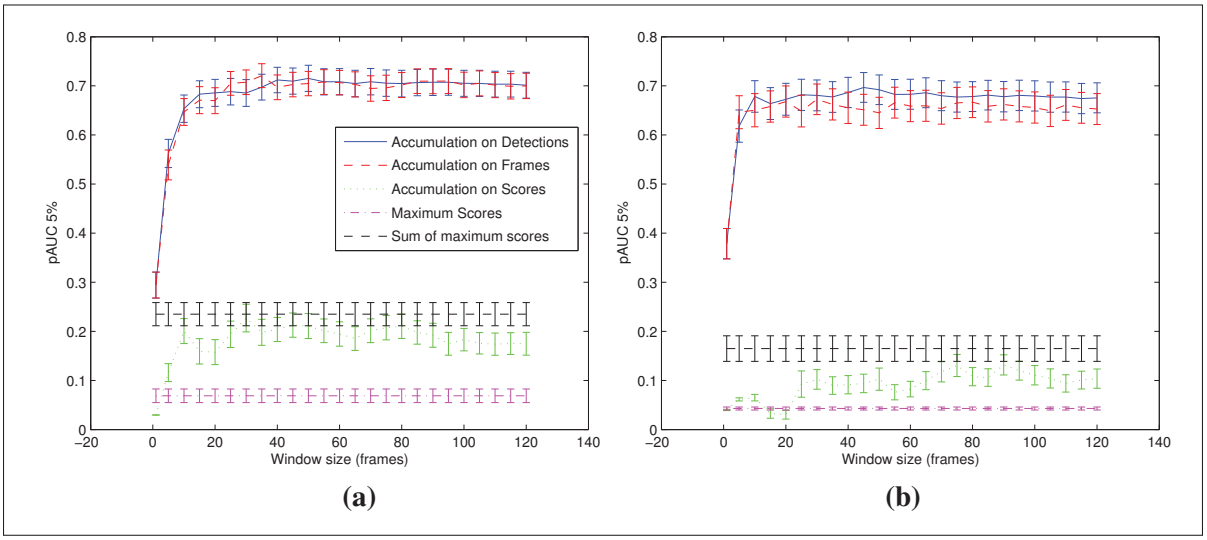


Figure 3.15 Impact of the window size on the pAUC (5%) produced by the system. The window size ranges from 0 to 4 seconds (1 to 120 frames), applied to the different fusion methods for the scenarios with (a) abrupt and (b) gradual changes

3.7 Conclusions

In this chapter, an adaptive ensemble-based system was proposed for spatio-temporal video-to-video FR, as found in person re-identification and search and retrieval applications. A pool of Probabilistic Fuzzy ARTMAP classifiers is generated using a DPSO learning strategy, and classifiers are selected and combined using Boolean combination. Classifiers are trained using the target samples from the trajectory, and a set of non-target samples selected from the cohort and universal models using One-Sided Selection. Each ensemble seeks to recognize target individuals and self-update facial models based on facial trajectories defined by the tracker, tunning up individual-specific parameters for classification and decision fusion, During operations, it integrates track IDs of a face tracker and predictions of a individual-specific ensemble at a decision-level for enhanced video-to-video FR. Classifier predictions for a trajectory are accumulated over a fixed-size time window window, and a detection threshold is applied for spatio-temporal fusion. A higher update threshold is applied to detect high confidence trajectories that serve for self-update. The set of facial captures linked to such trajectories for a target individual are used for self-update, to design an ensembles of 2-class classifiers or detectors (EoDs). A learn-and-combine strategy is then employed to avoid knowledge corruption during self-update of EoDs, and a memory management strategy based on Kullback-Leibler divergence is used to rank and select validation samples over time to avoid unbounded memory consumption.

The adaptive ensemble-based systems was validated with real-world Face in Action videos that feature abrupt (pose) and gradual (aging) patterns of changes. Experimental results indicate that the proposed system allows for improved overall performance after self-update with operational face trajectories. It was also observed that a decrease in the ambiguity of the ensemble has a negative impact in the performance of the system after self-update. Transaction-based analysis shows that the proposed system allows to increase the average pAUC (5%) accuracy in about 3% for the scenario with abrupt changes, and about 5% for the scenario with gradual changes. Subject-based analysis reveals the difficulties faced to recognize under different face poses, affecting most significantly the performance of lamb- and goat-like individuals.

A comparison between different spatio-temporal fusion approaches shows that the proposed scheme produces higher trajectory-based pAUC (5%) accuracy than other approaches, even for different window sizes.

Even though the system deals with imbalanced training data using a selection strategy, future work should consider the operational class imbalance to adjust classification parameters and achieve a performance closely related to the real environment. In order to maintain ensemble diversity, it would be interesting to explore different classifier generation strategies to provide more robust ensembles. Although the system allows to limit memory growth with the number of validation samples, a resource management strategy is still required to control the constant growth of the pool of classifiers with each self-update, and maintain a high level of performance. In that sense, time- and performance-based pruning techniques should be applied according to the individual-specific behavior. Finally, the system was characterized in environments with gradual and abrupt changes, but it would be interesting to analyze the performance of the system in an environment where multiple individuals are simultaneously present in scene.

CHAPTER 4

ADAPTIVE SKEW-SENSITIVE ENSEMBLES FOR FACE RECOGNITION IN VIDEO SURVEILLANCE

Miguel De-la-Torre[1,2], Eric Granger[1], Robert Sabourin[1], Dmitry O. Gorodnichy[3]

[1] Laboratoire d'imagerie de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montréal, Canada

[2] Centro Universitario de Los Valles, Universidad de Guadalajara, Ameca, México

[3] Science and Engineering Directorate, Canada Border Services Agency, Ottawa, Canada

Paper submitted to the journal "Pattern Recognition", from Elsevier, November 2014

ABSTRACT

Decision support systems for surveillance rely more and more on face recognition (FR) to detect target individuals of interest captured with video cameras. FR is a challenging problem in video surveillance due to variations in capture conditions, to camera interoperability, and to the limited representativeness of target facial models used for matching. Although adaptive classifier ensembles have been applied for robust face matching, it is often assumed that the proportions of faces captured for target and non-target individuals are balanced, known a priori, and do not change over time. Recently, some techniques have been proposed to adapt the fusion function of an ensemble according to class imbalance of the input data stream. For instance, Skew-Sensitive Boolean combination (SSBC) is a active approach that estimates target vs. non-target proportions periodically during operations using Hellinger distance, and adapts its ensemble fusion function to operational class imbalance. Beyond the challenges of estimating class imbalance, such techniques commonly generate diverse pools of classifiers by selecting balanced training data, limiting the potential diversity produced using the abundant non-target data. In this chapter, adaptive skew-sensitive ensembles are proposed to combine classifiers trained by selecting data with varying levels of imbalance and complexity, to sustain a high level of performance for video-to-video FR. Faces captured for each person in the scene are tracked and regrouped into trajectories. During enrollment, captures in a reference trajectory

are combined with selected non-target captures to generate a pool of 2-class classifiers using data with various levels of imbalance and complexity. During operations, the level of imbalance is periodically estimated from the input trajectories using the HDx quantification method, and pre-computed histogram representations of imbalanced data distributions. This approach allows to adapt pre-computed histograms and ensemble fusion functions based on the imbalance and complexity of operational data. Finally, the ensemble scores are accumulated of trajectories for robust spatio-temporal recognition. Results on synthetic data show that adapting the fusion function of ensemble trained with different complexities and levels of imbalance can significantly improve performance. Results on the Face in Action video data show that the proposed method can outperform reference techniques (including SSBC and meta-classification) in imbalanced video surveillance environments. Transaction-based analysis shows that performance is consistently higher across operational imbalances. Individual-specific analysis indicates that goat- and lamb-like individuals can benefit the most from adaptation to the operational imbalance. Finally, trajectory-based analysis shows that a video-to-video FR system based on the proposed approach can maintain, and even improve overall system discrimination.

4.1 Introduction

Video surveillance systems commonly rely on spatio-temporal face recognition (FR) to detect the presence of target individuals of interest in live or archived videos, either for watchlist screening or search and retrieval applications. Video-to-video FR systems commonly match input facial trajectories[1] from videos against the facial models of all target individuals enrolled to the system, and raise a warning in the case of positive detection. In this challenging scenario several persons may appear before a camera view point, and their appearance varies either abruptly or gradually due to, e.g., changes in illumination and pose. Changes in the capture conditions are associated with changes in the representation of the underlying class distribution of data in the face matching space. Uneven proportions between target and non-target individ-

[1] A trajectory is set of facial regions of interest (ROIs) captured in video that correspond to a same (high quality) track of a person appearing across consecutive frames.

uals are related to the prior probability of occurrence for a given individual, and are commonly referred to as class imbalance or skew.

Facial models used for matching are composed of a set of reference samples (for template matching), or a statistical model estimated during training with reference samples (for statistical or neural classification). For instance, some recent systems for face re-identification applications successfully employ adaptive ensembles of 2-class (target vs. non-target) classifiers to design and update facial models based on new reference trajectories, yet avoiding the knowledge corruption (De-la Torre *et al.*, 2012a, 2014a). And approaches to address the class imbalance problem in face recognition have also been proposed (Radtke *et al.*, 2013a,b). This chapter focuses on the design of facial models based on adaptive skew-sensitive ensembles of 2-class classifiers.

The effects of class imbalance on classifier performance have been shown by several authors (Guo *et al.*, 2008; Landgrebe *et al.*, 2006; Forman, 2006; Lopez *et al.*, 2013), and pattern recognition literature presents several ensemble-based methods to train ensembles on imbalanced data (Galar *et al.*, 2011). Algorithms designed for environments with data distributions that change over time can be categorized according to the use of a mechanism to detect concept drift or change (Ditzler and Polikar, 2013). Approaches with active detection of changes in prior probabilities seek explicitly to determine whether and when a change has occurred in the prior probability before taking a corrective action (Radtke *et al.*, 2013a,b; Ditzler and Polikar, 2013). Conversely, passive approaches assume that a change may occur at any time, or is continuously occurring, and hence classification systems are updated every time new data becomes available (Ditzler and Polikar, 2013; Oh *et al.*, 2011). The advantage of active approaches mainly consists in the avoidance of unnecessary updates. However, they are prone to both false positive and false negative drift detections, with the respective false updates and false no-updates. Passive approaches avoid these problems at an increased computational cost due to the constant update.

A representative example of active approaches for changing imbalances is the skew-sensitive Boolean combination (SSBC) that continuously estimates the class proportions using the Hellinger distance between histogram representations of operational and validation samples (Radtke *et al.*, 2013b). Every time the operational imbalance changes, SSBC selects one of the pre-calculated fusion functions that correspond to a set of prefixed imbalances. However, the limited number of validation imbalance levels that can be used to approximate the imbalance in operations is a limiting factor for the estimation of operational imbalance. Rather than selecting the closest imbalanced histogram representations, more sophisticated estimation methods may be employed for accurate estimation of the class proportions. Moreover, although it is scarcely exploited, the abundant non-target samples in video surveillance allow to produce training sets with different complexities and imbalances, and use them to generate diverse pools. A specialized combination and selection scheme of these diversified pools may lead to robust ensembles, considering both the different levels of complexity and imbalance (Lopez *et al.*, 2013).

In this chapter, adaptive skew-sensitive classifier ensembles are proposed for video surveillance applications. The proposed approach allows to select training data with varying levels of imbalance and complexity to design ensembles of classifiers that provide enhanced accuracy and robustness. Face captures of each person in the scene are tracked and regrouped into trajectories, and a decision threshold is applied to the accumulation of positive predictions from base classifiers for robust spatio-temporal recognition. During enrollment, facial captures from a reference trajectory are combined with selected non-target captures from the universal and cohort models[2] to generate a pool of 2-class classifiers using data with various levels of imbalance and complexity (class overlap and dispersion). Training/validation sets with different imbalances and complexities are built through random undersampling, and cover a range of imbalances from 1:1 to a maximum imbalanced estimated according to experience $1 : \lambda^{max}$. During operations, the operational level of imbalance is periodically estimated from the input data stream using the HDx quantification method, and pre-computed histogram representations

[2]The universal model (UM) is a database containing non-target trajectories from selected unknown people appearing in scene, and the cohort model (CM) is database with trajectories belonging to other target individuals enrolled to the system.

of imbalanced data distributions. The HDx quantification allows to estimate the prior probability of operational data based on the Hellinger distance between histogram representations of class distributions in the feature space, and employ a single validation set that is not required to provide a specific imbalance (Gonzalez-Castro *et al.*, 2013). Finally, the proposed approach allows to adapt pre-computed histograms and ensemble fusion functions based on the imbalance and complexity of operational data.

The proposed approach is validated with synthetic and video data, and compared against reference adaptive ensembles using BC, meta kNN fusion and score-level average fusion. The synthetic problem was designed to observe the impact of different theoretical probabilities of error as well as distinct imbalance levels in the performance of the system (Gaussian distributions in a two-dimensional feature space). The Carnegie Mellon University Face In Action (FIA) video database was used to emulate face re-identification applications. The transaction-based performance evaluates face matching of the system using the ROC and precision-recall spaces, and individual-specific characterization allows to analyze specific cases. Finally, trajectory-based analysis is employed to show the overall system performance over time.

The rest of this chapter is organized as follows. Section 4.2 presents a brief review of techniques for ensemble design (generation, selection and fusion) techniques, and specifically ensemble techniques proposed to address the problem of class imbalance. Section 4.3 describes the adaptive skew-sensitive ensembles proposed for FR in imbalanced environments. Section 4.4 provides synthetic experiments that motivated the proposed approach. Section 4.5 presents the experimental methodology and results with the FIA video data for validation of the proposed approach in face re-identification applications.

4.2 Ensemble Methods for Class Imbalance

Ensemble-learning techniques combine classifiers with diversity of opinions to increase classification performance. The design process can be divided into three main steps – generation of a pool of base classifiers, selection and fusion of classifiers (Duda *et al.*, 2001; Kuncheva, 2004;

Zenobi and Cunningham, 2001; Britto *et al.*, 2014). The first step allows to train base classifiers with diversity of opinions, and the last two take advantage of this diversity to produce more accurate predictions. Diversity can be created by employing distinct classifiers, train distinct instances of a classifier with different initial conditions (parameters), or using different training sets (Kuncheva, 2004).

Representative examples of ensemble methods are bagging, boosting, random subspaces, which employs different training sets of data or features from the training set to build distinct base classifiers (Kuncheva, 2004; Kittler, 1998). An example of diversity generation by various parameters is the work of Connolly et al. (Connolly *et al.*, 2012), which takes advantage of diversity in the hyperparameter space of classifiers to produce useful diversity of opinions. Examples of selection strategies are greedy search, clustering-based methods and ranking-based methods, and examples of fusion strategies can be divided in feature-based, score-based and decision-based (Tao and Veldhuis, 2008).

The algorithms designed for environments with changes in the probability distribution of data in general, and particularly in the class priors, can be categorized according to the use of a mechanism to detect changes in prior probabilities (Ditzler and Polikar, 2013). Approaches with active detection of changes in prior probabilities seek explicitly to determine whether and when a change has occurred in the prior probability before taking a corrective action (Radtke *et al.*, 2013a,b; Ditzler and Polikar, 2013). Conversely, approaches with passive change detection assume that a change may occur at any time, or is continuously occurring, and hence the classifiers are updated every time new data becomes available (Oh *et al.*, 2011; Ditzler and Polikar, 2013). The rest of this section describes representative approaches of passive and active ensembles for changing priors.

4.2.1 Passive Approaches

Passive *ensemble-based* methods for class imbalance can be categorized in cost-sensitive ensembles, boosting-based, bagging-based and hybrids (Galar *et al.*, 2011). In cost-sensitive ap-

proaches, the combination of classifiers (i.e. weights) is designed to consider the cost of class independent errors. Examples of these approaches include the AdaCost, CSB, RareBoost, AdaC1, AdaC2 and AdaC3 algorithms (Fan *et al.*, 1999; Wu, 2012). Boosting-based ensembles include techniques that use data preprocessing embedded into boosting algorithms. These methods bias the data distribution towards the minority class before the classifier generation step. Examples of these approaches are the Learn++.CDS, Learn++.NIE, SMOTEBoost, MSMOTEBoost, RUSBoost and DataBoost-IM algorithms (Ditzler and Polikar, 2013, 2010). Bagging-based ensembles integrate bagging with data preprocessing techniques, and hence, they do not require to update any kind of weights. These techniques address the class imbalance by the way they collect the training samples, using oversampling and/or undersampling techniques to generate training sets of different sizes. Examples of these techniques are the OverBagging, UnderBagging, UnderOverBagging and Imbalanced IVotes (Wang and Yao, 2009; Barandela *et al.*, 2003). Finally, hybrid ensembles combine a pre-processing technique with a bagging and a boosting technique. Techniques in this category are also called exploratory undersampling, and basically include EasyEnsemble and BalanceCascade (Liu *et al.*, 2009).

Although the aforementioned methods account for class imbalance through adaptation every time new reference samples become available, they are passive since they do not perform an estimation of the imbalance before adaptation. The advantage of passive approaches lies in the avoidance of false positive and false negative change detections, at the cost of the increased complexity of continuous adaptation.

4.2.2 Active Approaches

Active methods for adaptation to class imbalance employ a mechanism to estimate the class priors of the input data, and adapt the algorithm to the estimated class proportions when a change occurs. Hence, these approaches avoid the assumption of continuous changes and the complexity of continuous adaptations, with the potential disadvantage of false positive and false negative change detections. Several examples of active approaches that employ ensem-

bles for classification in imbalanced environments appear in literature (Radtke *et al.*, 2013a,b; Wang *et al.*, 2013a). In general, passive approaches for changing imbalance can be modified by adding a mechanism to detect changes in prior probabilities. Some examples of such mechanisms are based in Hellinger distance (Radtke *et al.*, 2013b), Kullback Leibler divergence (du Plessis and Sugiyama, 2012), or accounting for class-specific performance measures like *recall* (Wang *et al.*, 2013a,b).

A recently proposed active approach employed in face recognition in video surveillance is the skew-sensitive Boolean combination (SSBC), which estimates the imbalance using the Hellinger distance between the distributions of validation data and the most resent unlabeled operational samples (Radtke *et al.*, 2013b). During training, SSBC assumes that a diversified pool of binary classifiers $\mathscr{P} = \{p_1, ..., p_n\}$, and operates at the combination level to take advantage of the diversity of opinions in the ensemble. To do that, validation data with different levels of imbalance is used to estimate the operations points of the Boolean combination function (covering the whole ROC space). Two validation sets with that imbalances, the first (`OPT`) employed to estimate the operational imbalance, and the other (`VAL`) to select the operation point with the propper estimated imbalance. During operations, the imbalance is estimated using the Hellinger distance, and the operation points are selected from the predefined imbalances. The known levels of class imbalance used by the approach form the set $\Lambda = \{\lambda^{bal} = 1:1, ..., \lambda^{max}\}$. A subset of class imbalances $\Lambda_{BC} \subset \Lambda$ is selected from Λ to optimize a subset of BCs E. The subset of imbalances Λ_{BC} should contain evenly distributed intermediate class imbalance levels between the minimum λ^{bal} and the maximum level of imbalance λ^{max} inclusively. The sets `OPT` and `VAL` are generated from imbalanced reference data that follows λ^{max}. Different data sets with the levels of class imbalance defined in Λ, in which the amount of target samples remains fixed, while the amount of non-target samples are added to the set through random under sampling.

The classification system process streams of input patterns. The operational histogram `opd` corresponding to these operational samples is accumulated over time, and the closest level of class imbalance $\lambda^* \in \Lambda$ is estimated by comparing `opd` to the data sets in `OPT` using the

Hellinger distance. The estimated operational class imbalance λ^* corresponds to the imbalance of the closest set in OPT to opd in terms of Hellinger distance. Then, λ^* is used to select the BC that corresponds to that imbalance, and in the case λ^* is not available on Λ_{BC}, the BCs for the two closest imbalances are merged, and the convex hull is estimated.

The strength of the SSBC algorithm lies in the adaptive selection of suitable fusion functions (ROC operations points) according to the estimated operational imbalance. However, this technique assumes that the generation of a pool of classifiers, where each classifier is trained using balanced target and non-target data, and provide enough diversity of opinions to discriminate when input operational data is imbalanced. Another issue is related to the precision of the method used by SSBC to estimate the class imbalance is limited by the amount and sampling strategy used to create the set of imbalances Λ. Specialized methods to quantify the class priors of unlabeled (operational) data have been proposed in literature (Gonzalez-Castro *et al.*, 2013), and two of them are summarized in the next section.

4.2.3 Estimation of Class Imbalance

Quantification (i.e. estimation of the class distribution in Bayesian terms) is the task that deals with the estimation of the number of samples belonging to each class in an unlabeled set (Forman, 2006; Bella *et al.*, 2010; Forman, 2008). In the literature, different quantification methods appear and are based either on the classifier confusion matrix (Forman, 2006; Chan and Ng, 2006), the posterior probability estimates provided by a classifier (Bella *et al.*, 2010), or the comparison of class conditional probability densities of data sets with known and unknown proportions (Radtke *et al.*, 2013b; Gonzalez-Castro *et al.*, 2013; Forman, 2008; González-Castro *et al.*, 2010). Regarding the estimation task from the point of view of a classification algorithm, two levels can be identified to estimate the class imbalance of a distribution represented by a set of unlabeled (operational) samples. Data-level estimation operates in the feature space, employing the probability distribution of samples for each feature (Radtke *et al.*, 2013a,b; Gonzalez-Castro *et al.*, 2013). On the other hand, score-level allows to employ the probability distribution of the scores generated by a probabilistic classifier.

Two representative quantification methods were recently proposed to use the Hellinger distance to estimate the prior probability of unlabeled data, either using the features (HDx quantification) or scores from a classifier (HDy quantification) (Gonzalez-Castro *et al.*, 2013). Given an unlabeled dataset $U = \{(\mathbf{a}^n), n = 1, ..., N\}$ and a labeled validation dataset $V = \{(\mathbf{a}^m, l^m), m = 1, ..., M\}$, the Hellinger distance between these two sets can be computed according to

$$HD(V,U) = \frac{1}{n_f} \sum_{f=1}^{n_f} HD_f(V,U), \tag{4.1}$$

where the feature-specific Hellinger distance is given by

$$HD_f(V,U) = \sqrt{\sum_{i=1}^{b} \left(\sqrt{\frac{|V_{f,i}|}{|V|}} - \sqrt{\frac{|U_{f,i}|}{|U|}} \right)^2}, \tag{4.2}$$

where n_f is the number of features, b is the number of bins used to construct the feature-specific histogram representation of the probability density functions of the datasets. $|U|$ is the number of samples in U and $|U_{f,i}|$ is the number of samples whose feature f belongs to the bin i, similarly with $|V|$ and $|V_{f,i}|$ for the validation set V. The Hellinger distance between the probability densities of the unlabeled and validation sets can be computed by making the assumption

$$\frac{|V_{f,i}|}{|V|} = \frac{|S^-_{f,i}|}{|S^-|} P_v(-) + \frac{|S^+_{f,i}|}{|S^+|} P_v(+), \tag{4.3}$$

where $|S^-|$ is the number of non-target training samples and $|S^-_{f,i}|$ is the number of non-target samples whose feature f belongs to bin i in the histogram representation of the probability distribution of the training data S. Similarly, $|S^+|$ and $|S^+_{f,i}|$ are equivalent measures for the target class. The prior probability $P_v(+)$ (and similarly $P_v(-)$) can be manually assigned by the quantification method (see Algorithm 4.1). Algorithm 4.1 summarizes the process followed by the HDx quantification method.

Algorithm 4.1: Quantification HDx, extracted from (Gonzalez-Castro *et al.*, 2013)

Input : Labeled data S; operational data U (non-labeled); Number of bins b;
Output : Estimated target prior probability for U: $\hat{P}(+)$
Compute $|S^+|$, $|S^-|$ and $|U|$
for $f = 1...n_f$ **do**
  for $i = 1...b$ **do**
    Compute $|S^+_{f,i}|$, $|S^-_{f,i}|$ and $|U_{f,i}|$
for $P_v(+) = 0...1$ *in small steps* **do**
  for $f = 1...n_f$ **do**
    Compute HD_f according to (4.2), using (4.3) with $P_v(+)$
  $HD[P_v(+)] = \frac{1}{n_f}\sum_{f=1}^{n_f} HD_f[P_v(+)]$
$\hat{P}(+) = \arg\min(HD)$
$\hat{P}(-) = 1 - \hat{P}(+)$

For HDy, the Hellinger distance between the distributions of classifier outputs is estimated as

$$HD(V,U) = \sqrt{\sum_{i=1}^{b}\left(\sqrt{\frac{|V_{y,i}|}{|V|}} - \sqrt{\frac{|U_{y,i}|}{|U|}}\right)^2} \tag{4.4}$$

where $|U_{y,i}|$ and $|V_{y,i}|$ are the number of unlabeled and validation samples whose output y belongs to the bin $i = 1...b$. Similarly to the HDx method, the substitution to avoid subsampling and/or oversampling is given by

$$\frac{|V_{y,i}|}{|V|} = \frac{|S^-_{y,i}|}{|S^-|}P_v(-) + \frac{|S^+_{y,i}|}{|S^+|}P_v(+), \tag{4.5}$$

where $|S^+_{y,i}|$ and $|S^-_{y,i}|$ represent the number of non-target samples whose output y belongs to bin i in the histogram representation of the probability distribution of the scores. Algorithm 4.2 summarizes the process followed by HDy quantification to obtain the prior probability based.

4.2.4 Challenges

Exploiting imbalance to adapt a classifier system has been studied in literature, and is a consequent option regarding the imminent imbalance in face based video surveillance. Although the algorithms like SSBC have successfully used imbalanced validation data to update an en-

Algorithm 4.2: Quantification HDy, extracted from (Gonzalez-Castro *et al.*, 2013)

Input : Labeled data S; Operational data U (non-labeled); Classifier C_w; Number of bins b;
Output : Estimated target prior probability for U: $\hat{P}(+)$
Compute $|S^+|$, $|S^-|$ and $|U|$
Compute classifier outputs for S as $\{y^k = C_w(\mathbf{a}^k), k = 1,...,K\}$ Compute classifier outputs for U as $\{y^l = C_w(\mathbf{a}^l), l = 1,...,N\}$ **for** $i = 1...b$ **do**
  Compute $|S^+_{y,i}|$, $|S^-_{y,i}|$ and $|U_{y,i}|$
for $P_v(+) = 0...1$ *in small steps* **do**
  Compute $HD[P_v(+)]$ according to (4.2), using (4.3) with $P_v(+)$
$\hat{P}(+) = \arg\min(HD)$
$\hat{P}(-) = 1 - \hat{P}(+)$

semble fusion function to the operational imbalance, two issues are still to be addressed in practice. The first is related to the source of diversity of opinions among experts, where classifiers may be trained on data with different imbalances and complexities. In this way, the base classifiers trained on diverse levels of imbalance would provide increased useful diversity in the ensemble. Even more, training imbalance specific classifiers on data with different complexities would provide even more diversity, leading to a more accurate and robust ensemble under such an imbalanced environment.

The second issue is related to the resolution needed to reliably estimate the operational imbalance. For example, SSBC estimation relies on the measurement of the Hellinger distance between the histogram representation of a set with the most recent operational samples and validation sets with pre-defined imbalance levels (Λ). If the operational imbalance is not considered in the set Λ, the combination functions corresponding closest adjacent imbalances are considered, but the exact level of imbalance is never estimated. More accurate candidate quantification methods like HDx and HDy may be used, where all the validation samples are employed for a more precise estimation, avoiding the subsampling requirement. Moreover, the prior probabilities $P_v(+)$ and $P_v(-)$ are explicit – in other words, the step size in algorithms 4.1 and 4.2. The optimal size of each "small step" can be easily deducted by considering the maximum expected imbalance λ^{max}, which can be used to estimate the optimal size for these steps (See Section 4.4.2.4).

4.3 Adaptive Skew-Sensitive Ensembles for Video-to-Video Face Recognition

The proposed architecture for skew-sensitive video-to-video FR is depicted in Figure 4.1. It consists of a tracker, a skew-sensitive classification system with individual-specific parameters, a spatio-temporal fusion module, a sample selection and a classifier design/update systems. It is inspired on the framework proposed in (De-la Torre *et al.*, 2014a), and incorporates the functionality provided by skew-sensitive ensembles to adapt the individual-specific ensembles to the most recent operational imbalance. In order to adopt this functionality, some of the original blocks were modified, and others related to the operation skew-sensitive ensembles were added. The system works in two different phases that separate normal operation from the design and update of facial models of enrolled individuals.

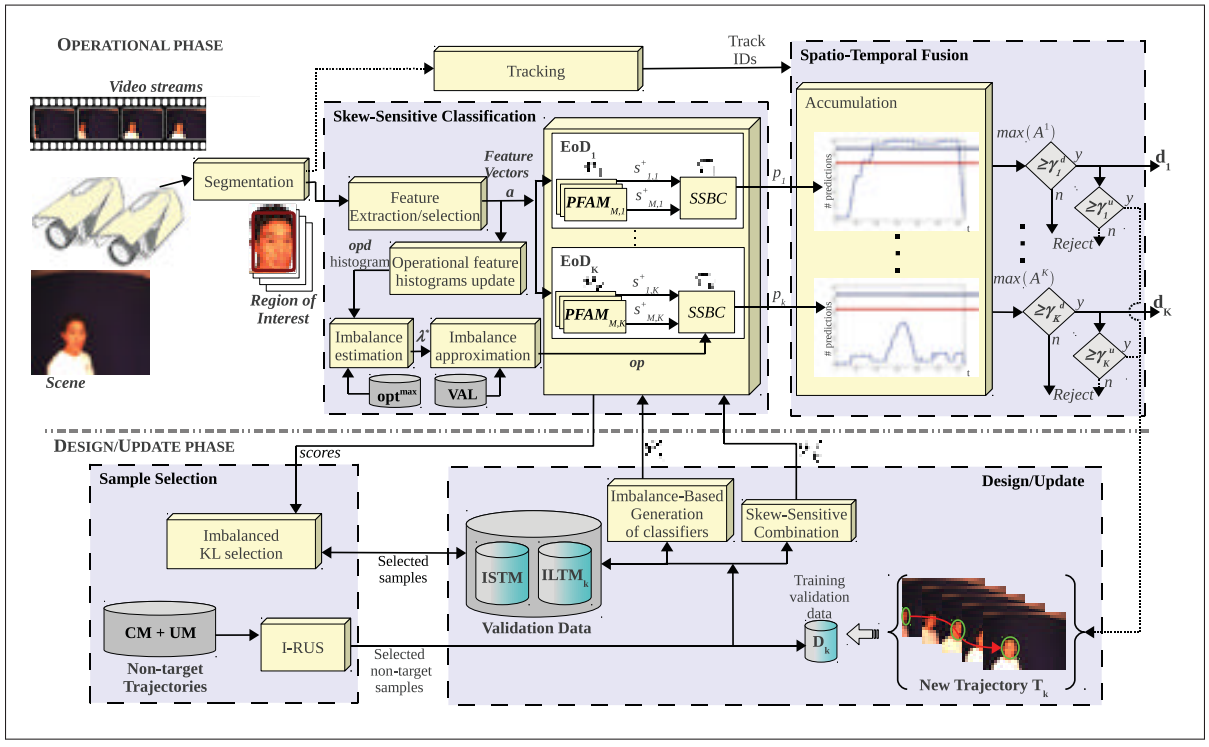


Figure 4.1 Adaptive skew-sensitive MCS for video-to-video FR

In the operational phase, the tracker follows the position of the segmented faces in video, building a face trajectory composed of sequential ROIs. Simultaneously, features for classification

are extracted and selected from segmented ROIs to form feature vectors (**a**), which are sequentially feed to all the individual-specific ensembles of classifiers. Each skew-sensitive ensemble k –corresponding to the enrolled individual k– produces a sequence of predictions according to the input order of the ROIs belonging to a face trajectory. In order to adapt the fusion function to the most recent operational imbalance, the feature specific histogram representation of the distribution of the operational data (**opd**) from facial captures of the last predefined time period (e.g. 15 minutes) is computed. The most recent distribution stored in **opd** is employed to estimate the operational imbalance λ^* (see Section 4.3.1). Then, the combination function corresponding to the estimated operational imbalance λ^* is approximated, and the operations point (**op**) in each individual-specific ensemble is selected. Finally, the spatio-temporal fusion module accumulates ensemble predictions over a fixed size window of face detections. When the accumulation of predictions from an individual-specific ensemble that corresponds to a trajectory surpasses a pre-defined detection threshold γ_k^d, the individual of interest k is detected in scene. For if self-update is required, the accumulation is compared to a second update threshold γ_k^u that triggers the adaptation process using all the ROIs belonging to the face trajectory (see (De-la Torre *et al.*, 2014a)).

The design/update phase is triggered when a new reference trajectory becomes available. Target samples are combined with non-target samples from UM and CM to form a learning data set D_k (for training and validation). The learning set D_k follows the maximum predefined imbalance λ^{max}, which is selected a priori in accordance with the experience in the field. An individual-specific selection strategy is employed to select the amount of non-target samples that accomplishes with the maximum expected imbalance λ^{max}. The learning data set D_k is evenly divided for imbalanced generation (D_k^{gen}) and validation of fusion functions (D_k^{val}). The imbalance-based generation of classifiers allows to generate a pool $\mathscr{P}'_k$ of classifiers, which are incorporated to the previous pool following a *learn-and-combine* strategy (see Section 4.3.2). A long term memory (LTM) is employed to store individual-specific reference samples and avoid knowledge corruption (De-la Torre *et al.*, 2013). Then, the validation samples used for combination are stored in the datasets $\circ pt^{max}$ for operational imbalance estimation (see Sec-

tion 4.3.1) and the approximation of imbalanced BC. Finally, the skew-sensitive combination allows to select the operations point with validation data with the approximated imbalance λ^* (see Section 4.2.2).

4.3.1 Approximation of Operational Imbalance

Initially, the classification system starts its operation considering a balanced operational environment. Feature vectors corresponding to input facial regions feed the data set with the most recent operational samples `ops`. The set `ops` is renewed with new input samples every certain prefixed period of time, let's say every 15 minutes. The operational feature histogram is estimated based on the evidence accumulated on the feature distributions of input facial regions during that period of time. Then, the prior probability of the most recent target class distribution $P^*(+)$ of operational samples is estimated using HDx quantization, based on the feature histograms from unlabeled operational (`ops`) and reference validation ($\texttt{opt}^{\texttt{max}}$) samples (Algorithm 4.1).

Let $|V^+|$ be the number of target samples in a validation data set V (e.g., $\texttt{opt}^{\texttt{max}}$). The number of non-target samples required to accomplish with the estimated class distribution $P^*(+)$ is given by

$$|V^-| = |V^+| \left(\frac{1}{P^*(+)} - 1 \right), \tag{4.6}$$

and the estimated class imbalance λ^* can be represented assuming $|V^+| = 1$ and substituting in the notation given by Eqn. 4.11.

The HDx quantification method require a single validation set ($\texttt{opt}^{max}$), which preserves the useful abundance of non-target samples that provide information from both imbalance and complexity in the feature space. The procedure for imbalance estimation is summarized in Algorithm 4.3.

The adaptation of the combination function for the new approximated class imbalance λ^* is performing in accordance to the skew-sensitive algorithm, either updating the combination

Algorithm 4.3: Estimation of the level of imbalance λ^* from reference data $\texttt{opt}^{max}$ and operational data ops

Input : Data set $\texttt{opt}^{max}$, Operational samples ops, number of bins b
Output : Imbalance estimation λ^*
Estimate prior probability using Algorithm 4.1
Assume $|V^+| = 1$
Compute $|V^-|$ using Eqn. 4.6
Compute imbalance λ^* using Eqn. 4.11

weights (weighted voting or meta-classification combiners) or selecting the imbalance-specific operations point (SSBC). The advantage of using an estimation of the prior probability as given by *HDx* provides a good estimation of the class imbalance, and the selection of the correct imbalance in validation set VAL reduces the error propagation induced by some algorithms for imbalance estimation (see Section 4.3.1).

4.3.2 Design and Adaptation of Ensembles

The imbalance-based generation strategy proposed in this section allows to generate useful diversity of opinions, which can be successfully exploited with other skew-sensitive combination strategies. The operational imbalance in a real scenario suffers from constant changes, and it is inaccurate to assume a single imbalance. Active skew-sensitive ensembles allow to estimate the operational imbalance, and select and combine the classifiers from a pool. Robustness of the ensembles may be enhanced with base classifiers trained on different levels of imbalance and complexity.

Limitations in resources make impractical to train a classifier for every possible imbalance, and a number of training imbalances should be fixed before training. The combination function is responsible for the selection of the classifiers with the proper imbalances according to the estimated operational imbalance. In this way, given predefined minimum and maximum imbalances denoted by λ^{min} and λ^{max} respectively, a fixed number of imbalances is chosen between them. Two issues appear from this affirmation, i.e. how to estimate the number imbalance levels are enough for the application, and how close should be from each other. The

first question is equivalent to estimate the number of classifiers in the ensemble that allow the fusion function to provide a high level of performance under distinct operational imbalances. The second question can be re-stated as which imbalances between the maximum and minimum should be used to train the base classifiers.

Algorithm 4.4: Generation of diversified classifiers based on different levels of imbalance and complexity

Input : Training data D_t, maximum imbalance λ_{GEN}^{max}, levels of imbalance $|\Lambda_{GEN}|$, size of subpools sp.

Output : $\mathscr{P}$ Pool of $|\Lambda_{GEN}| \times sp$ diversified classifiers.

Generate Λ_{GEN} by sampling the levels of imbalance with a log scale (e.g. 1:10^0, 1:$10^{\frac{log_{10}(\lambda^{max})\times i}{maxClsf-1}}$, ... 1:100)

Generate the imbalanced training sets D_i^{Imb} according to the imbalances in Λ_{GEN}

for i = *1*...$|\Lambda_{GEN}|$ **do**

  Train a new subpool with sp classifier $\mathscr{P}_i$ using D_i^{Imb} and a source of diversity

  $\mathscr{P} \Leftarrow \mathscr{P} \cup \mathscr{P}_i$

The proposed procedure for imbalance-based generation of diversified classifiers is shown in the Algorithm 4.4. In order to generate more diversity, the subpools of classifiers for each specific imbalance can be generated employing typical sources of diversity like different subsets of data, presentation orders, distinct hyperparameters, or other techniques (e.g. boosting, use different classification algorithms to train base classifiers, DPSO generation, etc.).

According to the results described in Section 4.4, $|\Lambda_{GEN}| = 7$ levels of imbalance are a good choice to train base classifiers, assuming that FR problems present high probability of classification error between target and non-target individuals. And the parameter that controls the size of the subpools may consider a small number of classifiers (e.g. $sp = 2$ or 3) to take advantage of complexity as a source of ensemble diversity, and train robust ensembles avoiding an excessive increase in memory requirements.

4.4 Synthetic Experiments

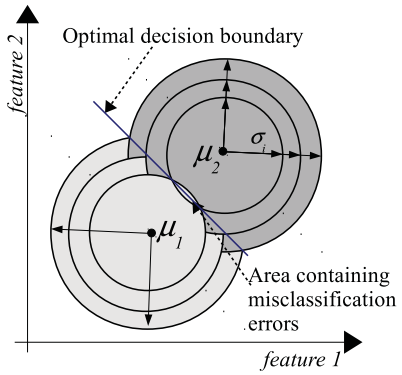
Consider a modular system used for matching in FR (Pagano *et al.*, 2012), where individual-specific ensembles of 2-class classifiers are trained independently. The scenario is replicated employing a Gaussian distribution to generate samples for the minority target class, and a second Gaussian distribution to draw samples for majority class (the rest of the individuals in the world).

The objective of these experiments is to characterize the performance of skew-sensitive ensembles with imbalance-generation of classifier ensembles in five axis. First, to show the capacity of the proposed imbalance-based generation of classifiers to produce ensemble diversity, since this affects positively the performance (accuracy and robustness) of ensembles. Second, a sensitivity analysis to decide the number of classifiers trained on different levels of imbalance levels that provide useful diversity to the ensemble. Third, to provide evidence of the effectiveness of the skew-sensitive ensembles in imbalanced environments compared to other ensemble techniques. Fourth, the generation of more than one classifier for each level of imbalance, bringing to the table the concept of imbalance-specific sub-pools. This approach provides combined sources of diversity from imbalanced training sets and different complexities. A sensitivity analysis allows to define size of the subpools that provide the best classification performance and robustness. And fifth, to provide a deep analysis of the behavior of the data- and score-levels employed in the approximation of imbalance employing quantification methods based on the Hellinger distance.

4.4.1 Experimental Protocol

The synthetic problem was designed in the 2 dimensional feature space, and the two overlapping multivariate Gaussian distributions with simple linear decision boundaries are shown in Figure 4.2a. Target and non-target data distributions are characterized by a fixed center of mass $\mu_1 = [0,0]$, $\mu_2 = [3.29,3.29]$ respectively, and the degree of overlap was variated by adjusting the covariance matrix σ of both distributions at the same time. The degree of overlap, and thus

the total probability of error between classes is variated according to six different levels, permitting the analysis of the impact of the overlap and imbalance level in the performance. The variances and levels of overlap of the class distributions used in these experiments are shown in Figure 4.2b. Ten different levels of imbalance were used to train 2-class PFAM clas-



(a) Synthetic overlapping

Total Probability of Error	Covariance Matrix
1%	$\mathbf{I}$
5%	$2 \times \mathbf{I}$
10%	$3.304 \times \mathbf{I}$
15%	$5.038 \times \mathbf{I}$
20%	$7.672 \times \mathbf{I}$
25%	$11.90 \times \mathbf{I}$

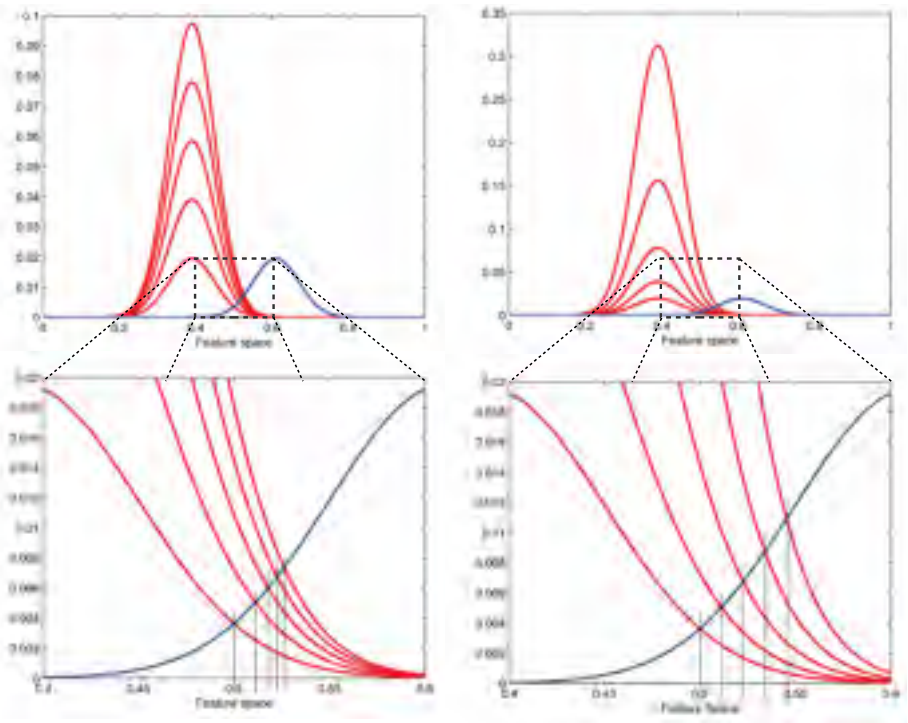
(b) Covariances matrices

Figure 4.2 (a) Representation of the synthetic overlapping data set used for simulations and (b) covariance matrices used to control the degree of overlap between distributions ($\mathbf{I}$ is the 2×2 identity matrix). The covariance matrix allows to change the degree of overlap, and thus the total probability of error between classes. These parameters were extracted from (Granger *et al.*, 2007)

sifiers[3], corresponding to a logarithmic sampling between balanced and the maximum level of imbalance $\lambda^{max} = 1:1000$. This sampling scheme was selected according to the following reasoning: First, it is recalled that the diversity of opinions between the base classifiers in an ensemble is an important characteristic for enhanced classification performance, and a good scheme should favor this diversity. Assuming no other source of diversity in an ensemble but the class imbalance, two similar classifiers (same algorithm and parameters) trained on data with different imbalance levels should produce different decision boundaries. Then, the scheme to subsample the space between λ^{1}_{GEN} and λ^{max}_{GEN} should maximize the diversity of

[3] $\Lambda_{GEN} = \{\lambda^{1}_{GEN}, ..., \lambda^{max}_{GEN}\} = \{1:1, 1:2, 1:5, 1:10, 1:22, 1:46, 1:100, 1:215, 1:464, 1:1000\}$

opinions, and hence produce distinct decision boundaries for each ensemble member. Figure 4.3 illustrates a linear and a logarithmic scheme, and the different theoretical optimal decision boundaries. It can be seen that a logarithmic scale produced a more even distribution of the decision boundaries along the feature space, thus generating a greater diversity of opinions between each classifier compared to the linear scheme. For this reason, the logarithmic scheme was chosen, allowing for enhanced diversity of opinions whereas evenly covering the space of decision boundaries for different imbalances. The standard hyperparameters of the PFAM



(a) Linear scheme **(b)** Logarithmic scheme

Figure 4.3 Cross-cut of the overlapping data distributions for target (right-blue curves) and non-target (left-red curves) samples. Linear scheme (a) with imbalances $\Lambda_{GEN} = \{1:1, 1:2, 1:3, 1:4, 1:5\}$ and logarithmic scheme (b) with imbalances $\Lambda_{GEN} = \{1:2^0, 1:2^1, 1:2^2, 1:2^3, 1:2^4\}$

classifiers were used (e.g. $[\alpha = 0.001, \beta = 1, \varepsilon = 0.001, \bar{\rho} = 0, r = 0.60]$), and a hold-out validation process was employed to optimize the number of training epochs with different orders

in the presentation of training samples. A constant number of 10 positive (target) samples was maintained in the training and validation sets, which is typical of applications where a limited amount of training samples is available. Similarly, the number of negative samples was variated according to the desired imbalances in Λ_{GEN}, assuming the wide availability of non-target samples, which is typical of surveillance scenarios where facial captures from non-target individuals may be easily retrieved from every day operational videos (the UM). The level of imbalance (prior probability) is internally estimated by the PFAM classifiers, based on the amount of positive and negative samples in the training data.

4.4.2 Results

4.4.2.1 Classification on Imbalanced Problems

Figure 4.4 shows the decision boundaries estimated by the ten classifiers trained on the imbalances in Λ_{GEN}. The test data set with 100 positives and the highest imbalance is plotted behind ($\lambda_{GEN}^{max} = 1:1000$), with blue for target and red for non-target samples. The differences in the decision boundaries is the agent that produces the diversity of opinions that can be exploited by ensemble techniques for increased robustness and accuracy.

The cost curves are graphical representations of the expected cost (or error rate) of 2-class classifiers over the full range of possible probability costs (class distributions or misclassification costs) (Drummond and Holte, 2006). In order to find the relation with the representations in the ROC and PROC spaces, the error rate can be defined as the difference between the false negative rate (fnr) and false positive rate (fpr) multiplied by the prior probability of a sample being from positive class $p(+)$ (see Eqn. 4.7). In Eqn. 4.7, the quantities of a 2-class confusion matrix are represented as FP or false positives, TP or true positives, FN or false negatives and FP or false positive predictions.

$$\begin{aligned} \textit{error rate} &= (FN - FP) * p(+) + FP \\ &= (1 - TP) * P(+) \end{aligned} \tag{4.7}$$

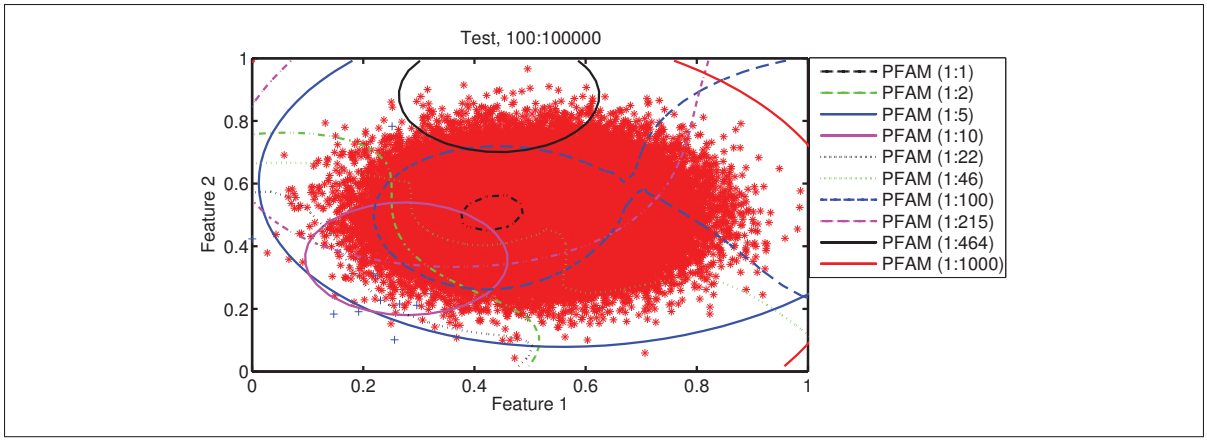


Figure 4.4 Test set characterized by a 1:1000 imbalance, and the decision lines drawn by the ten PFAM classifiers trained with different levels of imbalance in Λ_{GEN}. Classifiers and test samples correspond to the problem with a total probability of error corresponding to 20%

The extreme values of the x-axis in the cost curves represent the situations where all samples are classified as belonging to the same class. A point in the left extreme represents the probability of positives $p(x) = 0$ (all samples are negative), and a point in the right extreme represents $p(x) = 1$ (all samples are positive). Thus, a trivial classifiers can be represented with a cost curve that starts on the lower left corner (0, 0), grows linearly up to the point with equal positive and negative probabilities with error rate of 50% (0.5, 0.5), and ends at the lower right corner (1, 0). And a cost curve that corresponds to a perfect classifier should be drawn as a flat horizontal line at zero expected cost. On the other hand, the more commonly used receiver operating characteristics (ROC) curves plot the fpr in the x-axis and the tpr in the y-axis, with a point for each confusion matrix corresponding to an operational point (Fawcett, 2006). Similarly, the precision-recall (PROC) curves represent the *recall* (tpr) in the x-axis against the *precision* in the y-axis, although inverted axis can be employed to compare ROC and PROC spaces with the tpr in the y-axis. Examples of test ROC, PROC and cost curves(Drummond and Holte, 2006) for the ten PFAM classifiers trained with different levels of imbalance in the training set are shown in Fig. 4.5 for an overlap of 20%. The curves were obtained on a common test set with 100 positives and the highest imbalance used in the experiments (1:1000), the same as shown on Fig. 4.4. These results confirm that there is a significant difference in the

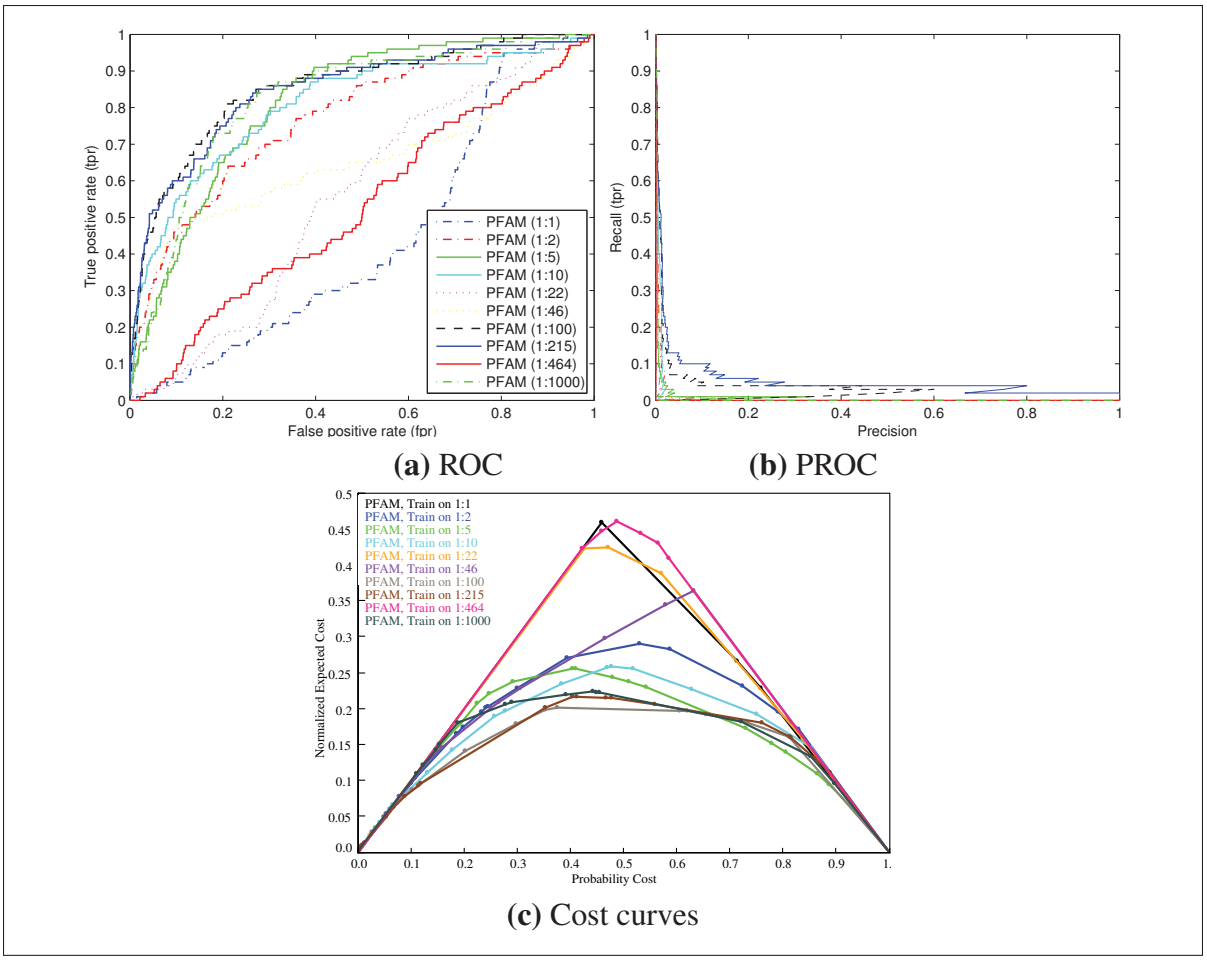


Figure 4.5 (a) ROC, (b) PROC and (c) cost curves corresponding to the seven PFAM classifiers trained on different imbalances, for the problem with a theoretical total probability error (overlap) of 20%

performance of the different classifiers, which constitutes a different view of the diversity of opinions provided by the distinct classifiers. This difference confirms the diversity of opinions that can be exploited using ensemble techniques, and that is related to the different levels of imbalance used in the training data.

In order to show the impact of training the classifiers with the same imbalance as the appearing in operations (test), each classifier was tested on a test set composed of 100 positive samples, and the necessary negative samples to complete the imbalance used for training. The experiment was repeated 10 times, and for each repetition the training data was randomly

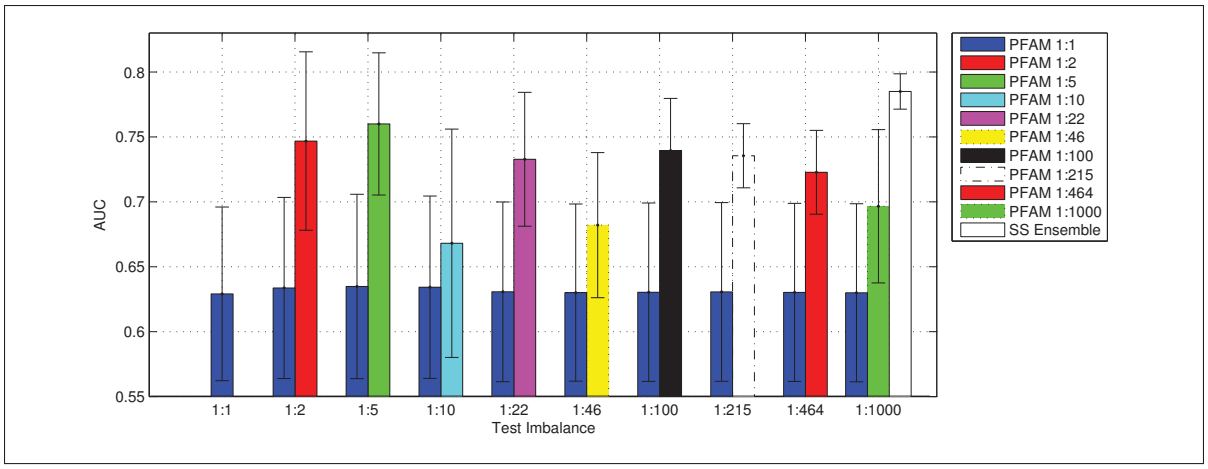


Figure 4.6 Average AUC estimated over 10 replications of the synthetic experiment with overlap between distributions of 20%. The left bar for each pair (blue) corresponds to the average AUC for the PFAM classifier trained on a balanced set (1:1), estimated on the test set with the imbalance indicated in the abscissa axis. The right bar for each pair is the average ROC AUC for the PFAM classifier trained on the same level of imbalance appearing in test

re-generated to design a completely independent experiment. After that, the ten classifiers were combined using skew-sensitive ensembles (SSBC), and the test set with the maximum imbalance ($\lambda_{GEN}^{max} = 1:1000$) was used to compare its performance with the single classifier approach.

Figure 4.6 presents the AUC performance for each of the classifiers and the skew-sensitive ensemble. It can be seen that the classifiers trained with the same level of imbalance that appears in test show a higher performance in terms of AUC than a classifier that learns from a balanced training set. Skew-sensitive ensembles estimate the level of imbalance in test and adapts the fusion function to the operational class proportions, providing the highest level of AUC performance and smaller standard error, as shown at the very right of the Fig. 4.6. A similar tendency was seen on the six levels of overlap, being more evident with higher probability of error. In general, as the probability of error increases, the problem is more difficult and the classifiers present lower performance, but the AUC performance of the ensemble was lower bounded by the performance of the best classifier in the ensemble.

4.4.2.2 Ensemble Generation

In order to define the levels of imbalance that provide the highest useful diversity to the imbalance-based ensemble, a sensitivity experiment was designed. The aim of this experiment is to explore how many of the ten classifiers are more useful for the ensemble, providing the best performance after selecting the operations point at a target $fpr = 1\%$. This scenario provides a situation where the ensemble is deployed and the operations point gives the final decisions, evaluating together the accuracy of the classifiers after the selection of the operations point, and not only a range of values in the ROC space. The number of PFAM classifiers was variated from 2 to 10, adding to the pool the classifiers in descendant order according to the ROC AUC accuracy evaluated on an independent validation set. In this way, the two classifiers that present the highest performance are first combined, then the third most accurate and so on, until the ensembles contain the ten classifiers trained according to the imbalances in Λ_{GEN}.

The five combination strategies used in the comparison are the max rule, average rule, meta kNN, BC and SSBC. The max rule selects the maximum target score produced by the base classifiers in the pool. The avg rule estimates the mean of the target scores produced by the base classifiers in the pool. In meta kNN, the 1NN classifier was trained on independent score-level validation data, and it was employed in test to produce output distance-based scores. In Boolean combination (BC), the ten Boolean functions are applied to different pairs of classifiers, and the BC algorithm was run on an independent validation set to find the operation points that maximize the ROC convex hull (Khreich *et al.*, 2012). Finally, the SSBC was applied with a validation set containing a profile with the same imbalance as the expected in test (Radtke *et al.*, 2013b).

In all cases, the operations point for a target $fpr = 1\%$ was selected using an independent validation set. The performance of all the approaches was evaluated on a same test set with imbalance $\lambda_{GEN}^{max} = 1:1000$, using precision and F_1 measure in the comparison, together with the ambiguity that measures the ensemble diversity. Formally, the ambiguity is defined by Zenobi and Cunningham in (Zenobi and Cunningham, 2001), and include the responses of the

base classifiers as well as the responses produced by the ensemble.

$$\text{Ens. Ambiguity} = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N} amb(\mathbf{a}_n, d_m, d^*), \tag{4.8}$$

where M is the number of classifiers in the ensemble, N is the amount of test samples. The ambiguity for an independent sample $\mathbf{a}_n$, given the decision d_m of the classifier c_m in the ensemble, is given by

$$amb(\mathbf{a}_n, d_m, d^*) = \begin{cases} 0 & \text{if } d_m = d^* \\ 1 & \text{otherwise.} \end{cases} \tag{4.9}$$

Figure 4.7 presents the resulting F_1 measure and ambiguity for the ensembles in the scenarios with total probability error of 15% and 20%. Regarding the F_1 measure, the maximum, average, BC and SSBC combinations perform better than meta-kNN at all times, and a significant superiority in performance is shown by SSBC when the ensemble contains between 5 and 8 classifiers. The phenomenon was repeated for the other overlaps (1%, 5%, 10% and 25%), becoming more evident as the total probability error grows. The ambiguity of the meta-kNN combination stays at a high compared to the other four approaches, which combined with the low performance shown in terms of F_1 measure, allows to see that this approach is the one that exploits the diversity of opinions in a less efficient way. On the other hand, the ambiguity shown by SSBC remains low compared to the meta-kNN, reinforcing that useful diversity of opinions is correctly exploited by this approach.

Regarding the F_1 measure in Fig. 4.7, it can be observed that the last value in the curve for SSBC in Fig. 4.7 (a), corresponding to 10 classifiers in the ensemble, is significantly higher than its starting point (2 classifiers). The same phenomenon was observed for the problems with total probability of error lower than 15%. However, in Fig. 4.7 (d) the same point in the curve presents an F_1 level that is only slightly higher than the starting point (2 classifiers). Similarly, this decrease in performance was observed in the problem with 25% total probability of error. This is related to the order used to add the base classifiers in the ensemble, in which the classifier with lowest level of performance is added in the last moment. This last classifier

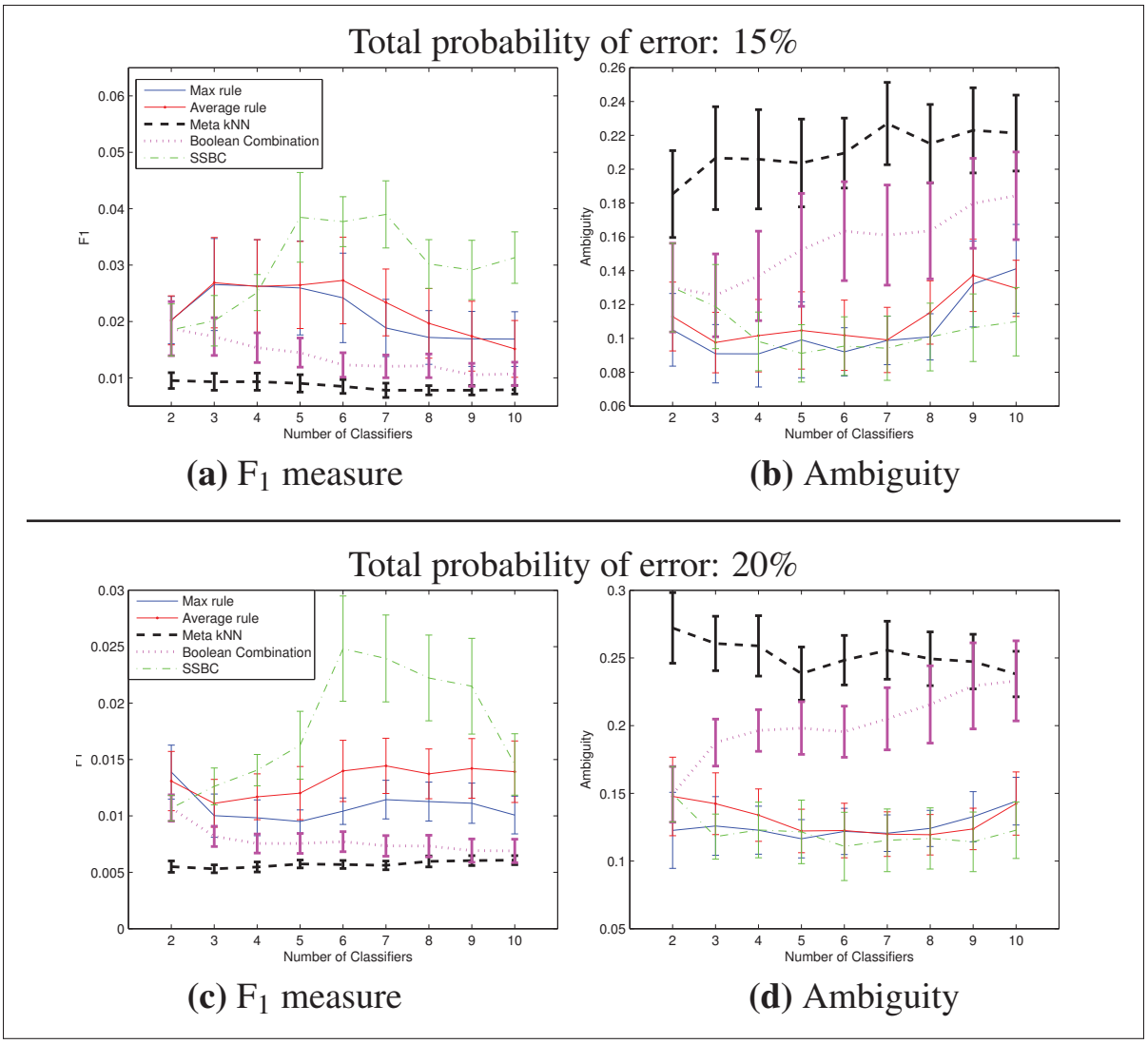


Figure 4.7 Sensitivity on the number of classifiers in the ensemble, using different combination strategies and adding the classifiers in descendant order according to the ROC-AUC evaluated on validation: the most accurate classifiers are the first added to the ensemble

negatively affects the diversity of opinions in the ensemble, and thus, the global performance. This tendency is more evident in problems with a high level of total probability of error, in which the classifiers with less performance bias the ensemble towards the erroneous decisions. And the classifiers with lower level of performance are commonly those trained with lower imbalance levels. For instance, regarding the problem with 20% total probability of error, in 8

of the 10 replications of the experiment, the classifier with less performance was trained with a training set with an imbalance lower than 1:50, and the imbalance used in test was 1:1000. In general, the approaches that show a higher diversity tend to produce a lower performance, showing that there is a limit in the useful diversity, and beyond that limit, it damages the ensemble accuracy.

Table 4.1 Average performance of the different combination methods, the ensembles are composed of 7 base classifiers. The bold numbers represent the performance values significantly higher than other approaches

Imbalanced PFAM				**Average**				**Meta kNN**				**SSBC**			
$fpr(\downarrow)$	$tpr(\uparrow)$	$prec(\uparrow)$	$F_1(\uparrow)$	$fpr(\downarrow)$	$tpr(\uparrow)$	$prec(\uparrow)$	$F_1(\uparrow)$	$fpr(\downarrow)$	$tpr(\uparrow)$	$prec(\uparrow)$	$F_1(\uparrow)$	$fpr(\downarrow)$	$tpr(\uparrow)$	$prec(\uparrow)$	$F_1(\uparrow)$
Total probability error: 1%															
13.26%	94.30%	1.61%	0.0314	11.74%	**99.90%**	1.57%	0.0307	1.37%	97.70%	**16.04%**	**0.2496**	**0.81%**	58.50%	6.85%	0.1219
(4.19)	(4.94)	(0.40)	(0.0078)	(3.06)	**(0.10)**	(0.39)	(0.0075)	(0.51)	(0.56)	**(5.03)**	**(0.0623)**	**(0.07)**	(5.53)	(0.45)	(0.0077)
Total probability error: 5%															
13.92%	50.30%	0.79%	0.0153	16.62%	**92.30%**	0.93%	0.0183	8.86%	87.40%	2.37%	0.0441	**0.93%**	57.50%	**6.17%**	**0.1102**
(3.70)	(11.06)	(0.32)	(0.0061)	(4.64)	**(2.21)**	(0.18)	(0.0034)	(2.16)	(2.33)	(0.99)	(0.0174)	**(0.08)**	(5.51)	**(0.73)**	**(0.0118)**
Total probability error: 10%															
12.32%	39.50%	0.75%	0.0140	13.71%	75.80%	1.40%	0.0267	15.67%	**81.70%**	0.62%	0.0122	**1.24%**	36.80%	**3.50%**	**0.0625**
(4.48)	(10.02)	(0.30)	(0.0054)	(4.16)	(5.32)	(0.47)	(0.0087)	(2.11)	**(3.80)**	(0.10)	(0.0019)	**(0.20)**	(4.07)	**(0.66)**	**(0.0106)**
Total probability error: 15%															
14.52%	42.00%	0.38%	0.0075	10.44%	49.10%	1.35%	0.0234	23.12%	**78.20%**	0.39%	0.0078	**1.13%**	21.80%	**2.16%**	**0.0390**
(3.55)	(9.58)	(0.10)	(0.0020)	(3.97)	(10.47)	(0.38)	(0.0059)	(2.50)	**(2.72)**	(0.06)	(0.0013)	**(0.13)**	(2.50)	**(0.34)**	**(0.0059)**
Total probability error: 20%															
19.00%	51.50%	0.28%	0.0057	11.99%	54.50%	0.74%	0.0144	27.88%	**75.00%**	0.28%	0.0056	**1.12%**	14.20%	**1.32%**	**0.0240**
(3.06)	(9.33)	(0.04)	(0.0007)	(3.77)	(5.68)	(0.13)	(0.0024)	(2.30)	**(2.56)**	(0.02)	(0.0004)	**(0.10)**	(2.13)	**(0.22)**	**(0.0038)**
Total probability error: 25%															
12.62%	32.40%	0.47%	0.0083	10.92%	42.20%	0.60%	0.0115	31.27%	**68.60%**	0.23%	0.0046	**1.22%**	8.10%	**0.67%**	**0.0123**
(3.78)	(6.52)	(0.15)	(0.0020)	(3.30)	(7.13)	(0.09)	(0.0017)	(2.56)	**(2.55)**	(0.02)	(0.0003)	**(0.10)**	(1.03)	**(0.07)**	**(0.0012)**

Table 4.1 allows for a more deep comparison between the different combination strategies, by considering 7 levels of imbalance in the ensembles. The empirical fpr and tpr were obtained after predictions for the selected operations point, together with the *precision* and F_1 measures. According to these results the SSBC provides the most accurate fpr in all cases, remaining always close to the desired $fpr = 1\%$ regardless of the total probability of error between classes. On the other hand, the average rule and meta kNN provide the highest tpr at the expenses of increased fpr, which is a costly trade off in video surveillance due to the amount of false alarms in an environment full of non-target individuals, or in other words, the operational imbalance. Comparing the F_1 measure for the different combination methods, it reflects that the SSBC significantly outperforms all other approaches, and is only the problem with an overlap of 1% that seems to be better addressed by the meta kNN. From this it can be

said that traditional combination methods are suitable to be used in imbalanced environments when the classification problems are easy enough –e.g. with lower total probability of error between classes, simple decision boundaries, etc. However, as the total probability of error grows, the superiority of the SSBC becomes more evident.

4.4.2.3 Using Several Classifiers per Imbalance

Up to here, a single classifier was trained for each imbalance level in Λ_{GEN}. However, using a single classifier per imbalance is not the only option to generate useful ensemble diversity and increase the robustness of the ensemble. Adding more than one classifier for each level of imbalance is a possibility that can be explored by generating a sub-pool instead of single classifiers. In this experiment, the number of classifiers in the ensembles was augmented by training more classifiers per imbalance, introducing variations in the classifiers by changing the presentation order in the training sets. A sensitivity analysis was conduced to observe the performance variations of the ensembles after changing the size of these sub-pools from 1 to 3 classifiers per imbalance, resulting in pools of 7, 14 and 21 classifiers. The test set was kept with the maximum imbalance ($\lambda_{GEN}^{max} = 1:1000$), and the samples were taken from the data distributions with 20% total probability of error.

Table 4.2 Average performance measures for the skew-sensitive ensemble with a pool of classifiers with 7 imbalances, problem with 20% total probability of error. A sub-pool for each of the imbalances was growth from one to three classifiers, resulting in pools of 7, 14 and 21 classifiers

	SS ensemble (7x1)	**SS ensemble (7x2)**	**SS ensemble (7x3)**
fpr$(\downarrow)$	1.12% (0.10)	0.97% (0.04)	0.96% (0.04)
tpr$(\uparrow)$	14.20% (2.13)	17.20% (1.16)	17.80% (1.01)
prec$(\uparrow)$	1.32% (0.22)	1.75% (0.11)	1.83% (0.09)
$\mathbf{F_1}(\uparrow)$	0.0240 (0.0038)	0.0317 (0.0020)	0.0331 (0.0017)

Table 4.2 shows the average performance of the skew-sensitive ensemble with 7 levels of imbalance using the three different sizes of sub-pools, at the operations point for $fpr = 1\%$. It can be seen that the best performance is achieved by the ensemble with 21 classifiers, at the expenses of an increased memory complexity, presenting the need to store three times more classifiers than using a single classifier per imbalance. The difference between using 7 and 14 classifiers is evident from the numbers in Table 4.2, showing that the ensemble with 14 classifiers presents a higher average performance and lower standard error. This is also true comparing the cases of 14 and 21 classifiers, but with a smaller difference in average performance and standard error. This confirms that more robust ensembles can be obtained by adding more classifiers to the sub pools, and the trade-off between resources and accuracy should be considered at the deployment stage.

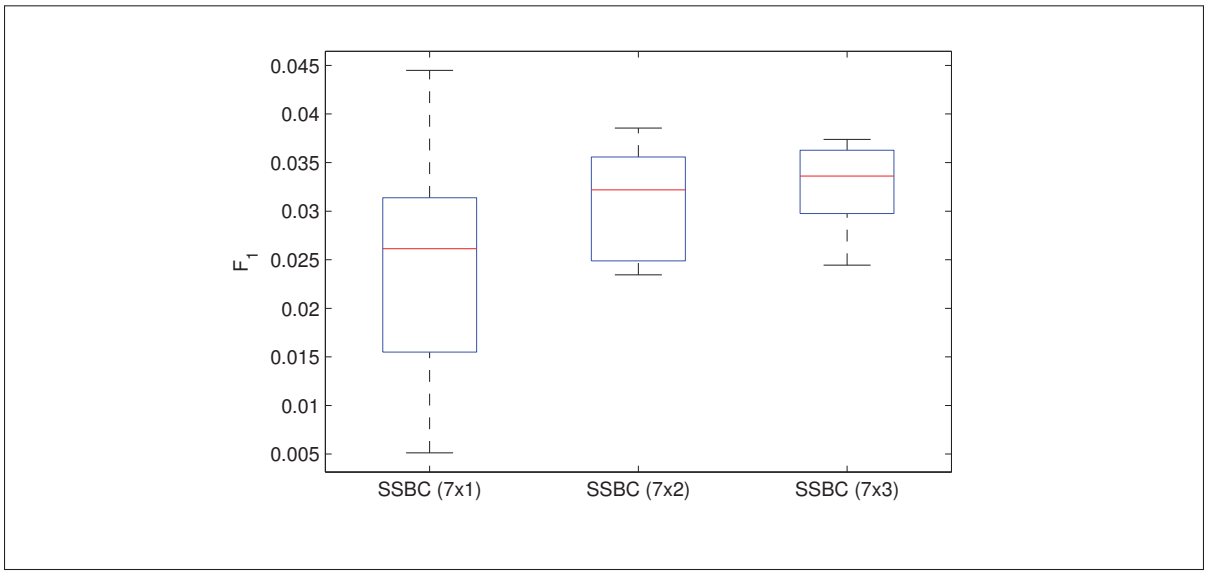


Figure 4.8 Box plots for the F_1 measure for the skew-sensitive ensemble with a pool of classifiers with 7 imbalances, problem with 20% total probability of error. A sub-pool for each of the imbalances was growth from one to three classifiers, resulting in pools of 7, 14 and 21 classifiers

Figure 4.8 presents the box plots for the F_1 measure achieved by the skew-sensitive ensemble with different sizes of pools of classifiers. It can be seen that the median is higher as the

number of classifiers increases, but there is also evidence of wide variations represented by the distance between upper and lower bars, which become narrower as the number of classifiers augments. The difference in the performance between the second (7x2 classifiers) and the third (7x3 classifiers) boxes is small, and other criteria like spatial complexity may be used to decide the size of the sub-pools.

4.4.2.4 Approximation of Imbalance Through Quantification

The level of class imbalance in the proportions of a set of samples is related to the prior probability of target (and equivalently non-target) samples. Given an imbalanced validation set V with $|V|$ samples, this relationship follows the definition of prior probability given by

$$P(+) = 1 - P(-) = \frac{|V^+|}{|V|} = \frac{|V^+|}{|V^+| + |V^-|}, \tag{4.10}$$

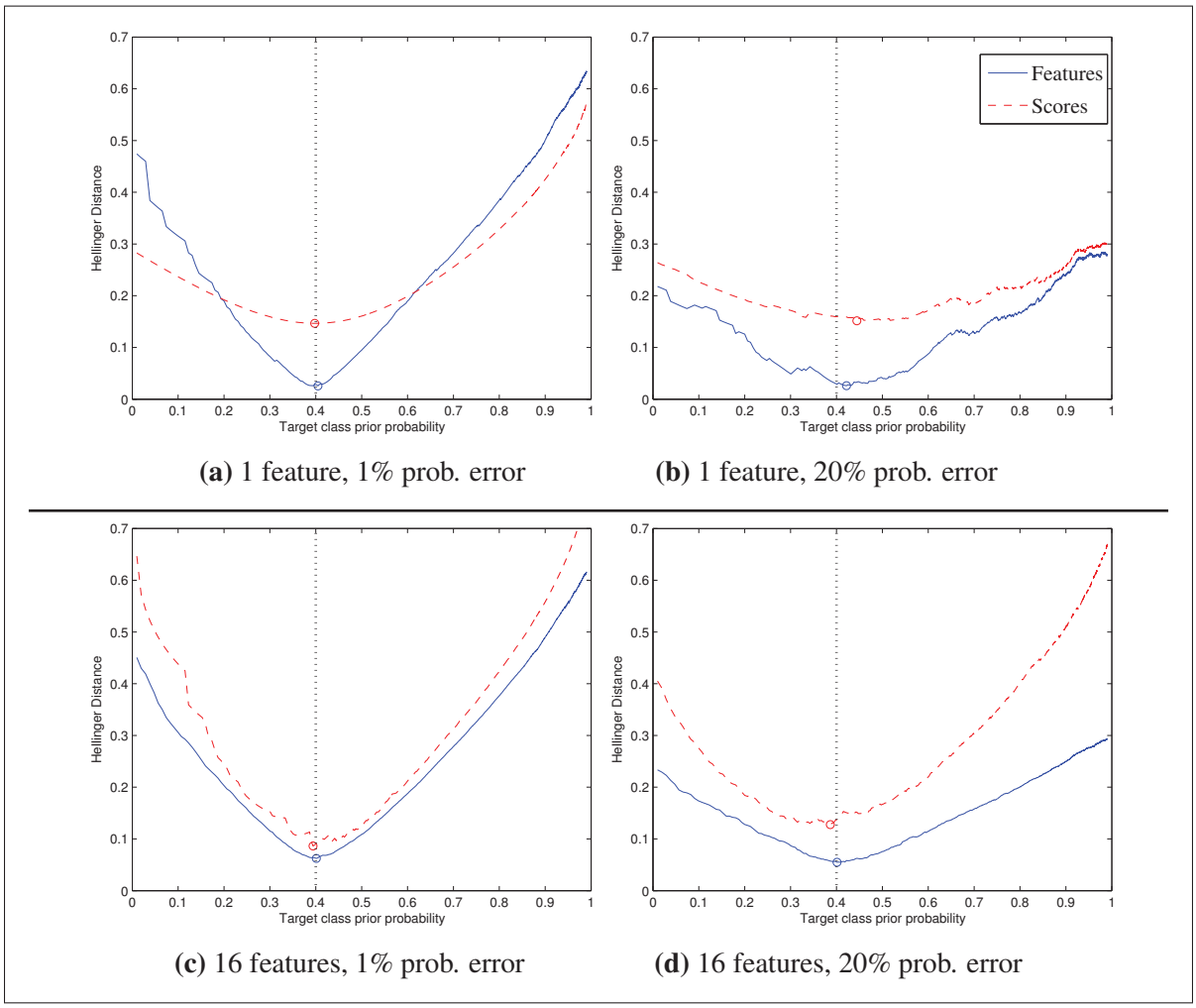
where $|V^+|$ and $|V^-|$ correspond to the number of target and non-target samples in V respectively. In the notation followed in this chapter, the level of imbalance is represented as

$$Imbalance = \frac{|V^+|}{|V^+|} : \frac{|V^-|}{|V^+|}, \tag{4.11}$$

and the number of target samples $|V^+|$ is given by the context. By simple algebraic substitution it is easy to see that both are representations of the same quantity. Hence, the HDx and HDy quantification methods provide an estimate of posterior probability $P(+)$, and equivalently, an estimate of the class imbalance.

The estimation of imbalance based on representations at feature (HDx) and score (HDy) spaces are characterized employing the Gaussian 2-class problem with different probabilities of error (see Figure 4.2b). The underlying probability densities employed to generate samples for the target ($P(x,+)$) and non-target ($P(x,-)$) were provided with prior probabilities $P(+) = 0.4$ and $P(-) = 0.6$ respectively. Binned distributions (histograms) for the test data were estimated after generating 1 000 samples for the joint distribution (400 target samples and 600 for non-target samples). Following this procedure, the original synthetic experiment with the

shift dataset was replicated (Gonzalez-Castro *et al.*, 2013), with the variant of a customizable overlap between distributions. The target prior probability $P(+)$ of the validation set kept a fixed amount of 100 target samples, whereas non-target samples were added one at the time to cover the different possible class prior probabilities. The probabilistic classifier employed to estimate the Hellinger distance at score level is the PFAM trained with balanced data.



(a) 1 feature, 1% prob. error **(b)** 1 feature, 20% prob. error

(c) 16 features, 1% prob. error **(d)** 16 features, 20% prob. error

Figure 4.9 HDx and HDy quantification examples related to the comparison between target and non-target distributions for the different cases.

The resulting Hellinger distance in feature and score spaces, corresponding to the low and high overlaps with single and multiple features are shown in Figure 4.9. In general, the HDy

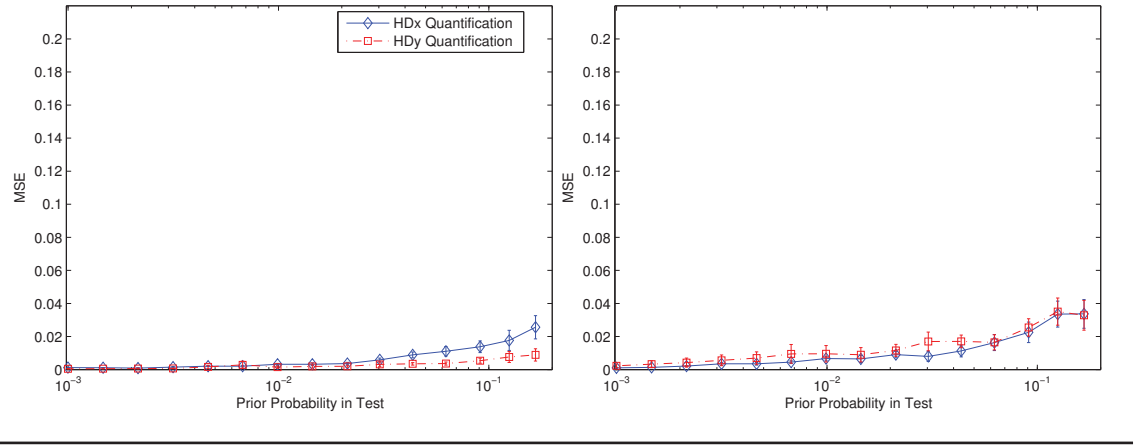
provides a softer curve for easy problems (e.g., small overlap between classes and few features). But as the standard deviation is increased and the overlap between probability densities grows, both curves present irregularities. Irregularities in the HDx curve are more evident with less features and a higher overlap between classes, but still using the distance based Hellinger distance provides a good estimation of the prior probability. Irregularities in the HDy curve increase with both the number of features and the overlap between classes, but still is capable of a good estimation of the prior probability. These irregularities in HDy are highly dependent on the complexity of the problem, and at the same time the accuracy of the classifiers employed to generate the estimated posterior probabilities (scores). Furthermore, the methods have been compared for a small fixed imbalance (1:2.5), but in video surveillance applications the imbalance is generally higher and changes over time.

The accuracy of the quantification methods was evaluated using data sets with 15 levels of imbalance, including 7 levels distinct to those used for training and validation. The samples were drawn from the overlapping Gaussian distributions described in this section. Test imbalances that appear in Λ are {1:5, 1:7, 1:10, 1:15, 1:22, 1:32, 1:46, 1:68, 1:100, 1:147, 1:215, 1:316, 1:464, 1:681, 1:1000}. Equivalently, the target prior probabilities of these datasets can be computed as $\{ \frac{1}{5+1} = 0.1667, \frac{1}{7+1} = 0.1250, ...\}$. A single validation set with the maximum level of imbalance was used with the quantification methods, avoiding the requirement of using several validation sets with different levels of imbalance. The size of the "small steps" in Algorithms 4.1 and 4.2 is set in accordance to the minimum possible probability, or equivalently, the maximum expected imbalance $\lambda^{max} = 1:1000$. The STEPSIZE employed in experiments was defined using the validation set V, and is given by

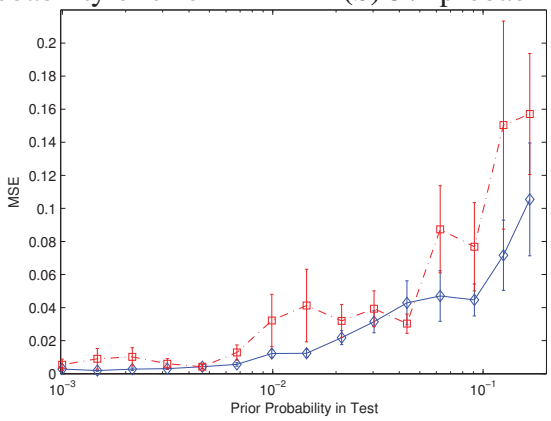
$$STEPSIZE = P_{min}(+) = \frac{V^+}{V^+ + V^-} \tag{4.12}$$

The average mean squared error between true prior probabilities and the estimations obtained with the HDx and HDy classification methods are shown in Figure 4.10. Comparing Figs. 4.10 (a), 4.10 (b) and 4.10 (c), it can be seen that the HDy quantification outperforms HDx

when the total probability of error is small, and HDx outperforms HDy as the classifiers are less accurate. This is consistent with the affirmation that HDy is more reliable when classifiers are more accurate, as stated in (Gonzalez-Castro *et al.*, 2013). We can reformulate and affirm that according to the results shown in Fig. 4.10, HDy is more reliable when the target and non-target samples are easily separable, but HDx is preferable for problems with higher total probability of error (e.g. overlap between class distributions).



(a) 1% probability of error **(b)** 5% probability of error



(c) 20% probability of error

Figure 4.10 Average mean squared error (MSE) between the true prior probability in test and the estimation produced using the quantification methods HDx and HDy

According to the observations in this section, the estimation of the class imbalance should be guided by the characteristics of the data employed in the application and the particular algorithm used for classification. In this chapter, an experiment was conduced to select the proper method, and the results are analyzed in Section 4.5.4.

4.4.3 Discussion

In conclusion, the following affirmations should be considered in the choice of parameters for systems that will operate in environments with changing class imbalance. First, the design of classifiers considering the imbalance expected in test allows the classifiers to outperform those classifiers that are trained with balanced data. Second, according to the simulations, a generation strategy that considers 7 imbalances to train base classifiers is a good choice, specially for problems that present high probability of error between classes. In general, the approaches that present a higher diversity tend to produce a lower performance, showing that there is a limit in the useful diversity, and beyond that limit it damages the ensemble accuracy. Third, from the combination methods analyzed in this section, skew-sensitive ensembles provide the highest level of performance in terms of F_1 measure in environments with different levels of imbalance. Fourth, the use of several classifiers per imbalance is an option to increase the performance of the ensemble and reduce the standard error, and the advantage has to be contrasted with the significant increase of the pool size at deployment time. Finally, quantification methods may be used within skew-sensitive approaches to obtain a more precise estimation of the operational imbalance, and HDx quantification is a good candidate specially when the total probability of error is high.

4.5 Experiments on Video Data

4.5.1 Experimental Protocol

This section presents the methodology used in simulations, following a video surveillance scenario using real data to demonstrate the effectiveness of the proposed imbalance-based

generation method, and the characterization of this method when combined with SSBC. The video-based FR system that was used as a model in the experiments is depicted in Figure 4.11. A single IP camera continuously captures the scene and feeds the segmentation module that isolates the facial regions of interest (ROIs) in each consecutive frame. After a first ROI is captured from an individual in scene, the tracking and classification modules are triggered in parallel. The tracking module starts following the individual's face and regrouping ROIs from a same individual in trajectories, whereas the classification module produces consecutive identity predictions for each ROI. Finally, the spatio-temporal decision fusion module allows to accumulate target predictions, and applies individual-specific thresholds for enhanced spatio-temporal FR, as described in (De-la Torre *et al.*, 2014a).

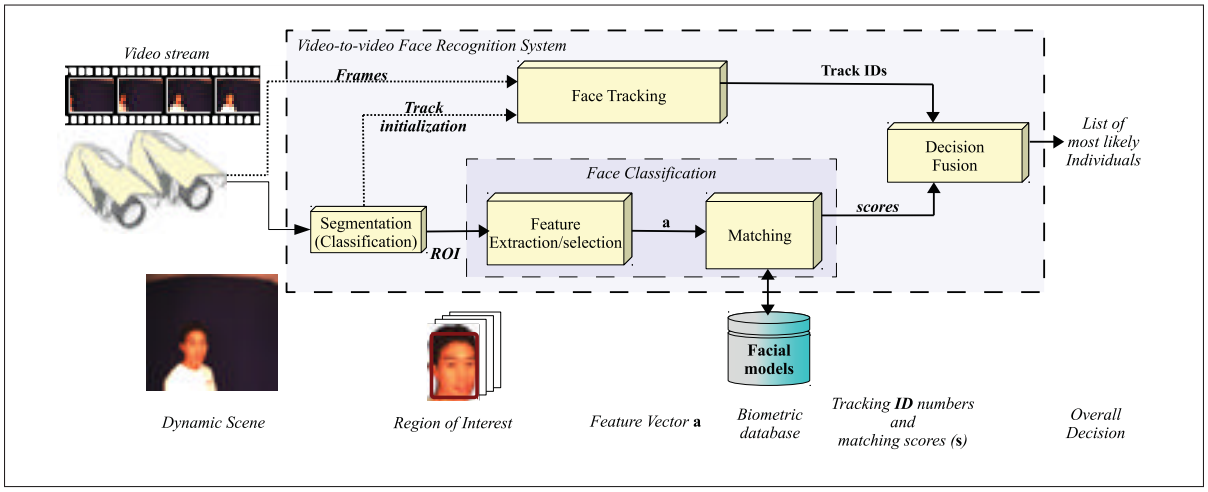


Figure 4.11 Generic video-based FR system used in video surveillance applications

In this particular implementation, the popular Viola-Jones face detector was used to extract grayscale ROIs (Viola and Jones, 2004). Pixel intensities are concatenated with multi-block local binary patterns (MBLBP) features, and the 32 principal components are selected after application of PCA. Training feature vectors **a** are used to design the biometric database, and the pixels of never seen ROIs are projected to the 32 dimensional feature space employed for face matching. Face tracking was implemented using the incremental visual tracking (IVT) algorithm, which incrementally learns the low-dimensional subspace representation (Eigen ba-

sis) by efficiently adapting online to changes in the appearance of the face model (Ross *et al.*, 2008).

The classification architecture used for matching is composed of an ensemble of 2-class ARTMAP classifiers for each individual. This architecture has been widely used for face matching in literature, which models the general recognition problem in terms of individual-specific detection problems (De-la Torre *et al.*, 2012b; Pagano *et al.*, 2012; Radtke *et al.*, 2013b). In the reference system used in comparison, the individual-specific EoDs are co-jointly trained using a DPSO learning strategy, which allows for the generation of a diversified pool of Probabilistic Fuzzy ARTMAP classifiers. The proposed approach preserves the same architecture, but the base classifiers are trained independently on different imbalances, using DPSO to optimize the hyperparameters and the global best is added to the pool.

4.5.2 Video Surveillance Data

Videos from the Carnegie Mellon University - Face in Action (FIA) database are used in experiments (Goh *et al.*, 2005). These videos correspond to 20 seconds sequences for 244 individuals that act simulating a passport checking scenario. Six cameras capture the scene at a resolution of 640×480 pixels, at a frame rate of 30 frames per second. Data was captured over three different capture sessions separated by a three months interval. The six cameras were distributed in three pairs with focal lengths of 2.8 mm (unzoomed) and 4.8 mm (zoomed), and positioned in horizontal positions with frontal, left and right orientations corresponding to 0^o, and $\pm 72^o$. In the experiment, a video stream from a single IP camera is formed using the frontal zoomed and unzoomed cameras along the three capture sessions.

Ten individuals of interest were selected from the FIA database for enrollment (FIA IDs 2, 58, 72, 92, 147, 151, 176, 188, 190 and 209), and the rest was divided into two independent subsets of non-target classes appearing in training and test. For each individual of interest, 100 non-target individuals are selected for training (UM and CM), and 100 different individuals are selected for test, providing a maximum class imbalance of $\lambda^{max} = 1:100$. The cohort and

universal models (CM and UM) allow to train 2-class ensembles with improved discrimination between target and non-target classes, training one EoD for each individual of interest as the target class, as described in (Pagano *et al.*, 2012).

4.5.3 Experimental Protocol

For enrollment, an adaptive skew-sensitive ensemble of classifiers was trained for each one of the selected individuals of interest. In the initial step, a pool of PFAM classifiers (Lim and Harrison, 1995, 1997) was generated using seven different imbalances for training. The DPSO learning strategy was used to co-jointly optimize the hyperparameters of a PFAM neural network for each imbalance in Λ_{GEN}, using training and validation data that follows the corresponding imbalances in Λ_{GEN}. The DPSO algorithm was initialized with a population size of 20 particles, a maximum of 6 subswarms of 5 particles maximum, and a maximum of 10 iterations (Granger *et al.*, 2007). At the end of the DPSO learning process, the global best classifier was selected as the classifier the level of imbalance that corresponds to each the training levels in Λ_{GEN}.

Let λ^1 and λ^{max} be the minimum and maximum possible imbalances in the classification environment. λ^{max} can be manually set according to the amount of detectable faces that can fit in a frame captured by the camera. The range of possible imbalances has to be sampled in as many imbalances as classifiers are required in the pool. Having established a maximum imbalance of $\lambda^{max} = 1:100$, five subdivisions were established in a logarithmic scale in order to obtain seven different imbalances between $\lambda^1 = 1:1$ and $\lambda^{max} = 1:100$. The resulting imbalances used are $\Lambda = \{1:1, 1:10^{1/3}, 1:10^{2/3}, 1:10, 1:10^{4/3}, 1:10^{5/3}, 1:100\}$.

Learning is performed following a 4×6-fold cross-validation process for 24 independent trials. Positive samples from the incoming sequence are randomly split according to a uniform distribution, in 6 folds of the same size. The first two folds combined in a training set (D^t), and the rest of the folds are distributed in validation sets used to stop training epochs (D^e), fitness evaluation (D^f), estimation of combination points in the ROC space (D^c) and selection of the

operational point (D^s). Four imbalance levels are produced for each training and validation set, picking different number of negative samples from the CM and UM. The levels of class imbalance used for the different test blocks are 1/20, 1/35, 1/80, 1/55, 1/100, 1/70, 1/50 and 1/15, for t=1, ..., 8 respectively. The changes in the class imbalance of the test sets are obtained by randomly removing individuals from each block of 30-min.

The proposed system is evaluated at transaction-level using the ROC and PROC spaces after selecting the operations point for a fixed $fpr = 1\%$. The operational measures used in the characterization are the fpr, tpr (or *recall*), *precision* and the F_1 measure. The ambiguity is used as a measure of the diversity of opinions generated by the base classifiers trained on different imbalances (Zenobi and Cunningham, 2001). Individual specific analysis is employed

Table 4.3 Doddington's zoo taxonomy for binary decisions. False negative rate (fnr) and false positive rate (fpr) thresholds are applied to each individual-specific ensemble

Category	**Target class**	**Non-target class**
Sheep	$tpr \geq 55\%$ and not a lamb	$fpr \leq 1\%$
Lamb	At least 5% of non-target individuals are wolves	-
Goat	$tpr < 55\%$ and not a lamb	-
Wolf	-	$fpr > 1\%$

following the Doddington's Zoo taxonomy (Doddington *et al.*, 1998; Rattani *et al.*, 2009a), with the thresholds shown in Table 4.3. Finally, time-based analysis is employed to see the adaptation of the system to the operational class imbalance over time.

4.5.4 Results

This section presents the results obtained after computer simulations, divided into four different levels of analysis. The first level presents transaction-based analysis, which corresponds to the evaluation of the classification system after the presentation of each single facial region, and its evolution as the system adapts to the imbalance in the environment. It is known that

biometric systems have different performance depending on each specific case, and the second level of analysis presents the individual-specific characterization of the system. The third level of analysis is related to the functionality of the system to perform operational imbalance estimation. Finally, as a video-based FR system, the trajectory level analysis presents the overall evaluation of the system for trajectories from the different individuals of interest.

4.5.4.1 Transaction-Based Analysis

Table 4.4 shows the average performance of the system for the different approaches after selecting the operations point for a desired $fpr = 1\%$. The first two approaches are the reference systems that use the baseline balanced DPSO generation method, either using BC or the proposed approach. These two approaches present the same initial performance, the F_1 score presented by the proposed approach after adaptation to each block of test data is higher than DPSO+BC. This superiority is product of the better estimation of the operations point when the fusion function considers the class imbalance in the environment, which results in a more accurate combination than employing balanced training without imbalance estimation. It is remarkable to observe that the proposed approach preserves a fpr closer to the desired $fpr = 1\%$, which is an evidence of the correct exploitation of the imbalance to select a more accurate operations point.

The last two approaches presented in Table 4.4 correspond to the same approaches as the first two, but replacing the balanced generation by the proposed imbalanced generation scheme. A similar trend can be observed when the combination methods are compared. The proposed approach overcomes the performance of imbalanced training + BC in terms of F_1 score, and the rejection of false positives (fpr) obtained by the proposed approach is more accurate than using imbalanced training + BC. This trend confirms that the adaptive capacity of the proposed approach provides a powerful tool for combination in environments with changing imbalance, regardless of the generation method.

Table 4.4 Average performance for different approaches for a target 1% fpr on test blocks at different t times, including the different individuals enrolled to the system. The standard error is detailed between parenthesis

Approach	**M**easure	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
Balanced training+BC	fpr	4.80% (0.032)	4.14% (0.023)	5.93% (0.030)	5.57% (0.023)	4.35% (0.024)	4.11% (0.022)	3.00% (0.014)	3.19% (0.021)
	tpr *recall*	57.02% (0.317)	57.63% (0.327)	58.39% (0.213)	59.49% (0.230)	58.09% (0.223)	56.29% (0.262)	55.61% (0.349)	54.70% (0.342)
	precision	43.28% (0.190)	36.82% (0.191)	20.09% (0.086)	26.04% (0.082)	24.11% (0.132)	27.47% (0.143)	36.99% (0.194)	54.96% (0.249)
	F_1	0.436 (0.225)	0.400 (0.226)	0.267 (0.110)	0.328 (0.117)	0.302 (0.155)	0.326 (0.172)	0.394 (0.248)	0.479 (0.284)
Balanced training+SSBC	fpr	4.80% (0.032)	1.17% (0.011)	1.61% (0.007)	1.69% (0.009)	1.17% (0.007)	1.08% (0.009)	0.55% (0.005)	0.62% (0.006)
	tpr *recall*	57.02% (0.317)	43.45% (0.293)	42.35% (0.231)	42.92% (0.257)	41.56% (0.281)	38.34% (0.311)	43.09% (0.343)	42.19% (0.335)
	precision	43.28% (0.190)	56.34% (0.300)	33.81% (0.163)	39.20% (0.144)	34.59% (0.184)	39.21% (0.208)	54.97% (0.303)	67.62% (0.313)
	F_1	0.436 (0.225)	0.428 (0.272)	0.339 (0.154)	0.372 (0.179)	0.339 (0.209)	0.338 (0.233)	0.441 (0.311)	0.453 (0.328)
Imbalanced training+BC	fpr	4.96% (0.037)	4.25% (0.025)	5.18% (0.025)	5.06% (0.021)	4.30% (0.025)	4.15% (0.022)	3.33% (0.015)	4.03% (0.028)
	tpr *recall*	59.78% (0.279)	60.09% (0.271)	61.32% (0.174)	59.97% (0.206)	59.36% (0.174)	57.12% (0.224)	59.42% (0.309)	59.67% (0.310)
	precision	44.11% (0.196)	39.17% (0.180)	23.72% (0.079)	28.59% (0.086)	23.19% (0.089)	26.11% (0.100)	37.45% (0.183)	54.82% (0.239)
	F_1	**0.456 (0.220)**	0.420 (0.209)	0.302 (0.094)	0.351 (0.115)	0.297 (0.111)	0.320 (0.131)	0.408 (0.220)	0.502 (0.261)
Proposed approach	fpr	4.96% (0.037)	1.78% (0.018)	1.69% (0.009)	1.92% (0.013)	1.52% (0.008)	1.49% (0.010)	1.06% (0.006)	1.60% (0.013)
	tpr *recall*	59.78% (0.279)	52.91% (0.290)	56.71% (0.192)	55.64% (0.281)	58.33% (0.270)	53.87% (0.348)	54.83% (0.364)	54.56% (0.360)
	precision	44.11% (0.196)	57.30% (0.302)	42.92% (0.155)	47.83% (0.182)	38.46% (0.136)	41.80% (0.166)	52.95% (0.263)	64.75% (0.315)
	F_1	**0.456 (0.220)**	**0.491 (0.262)**	**0.445 (0.112)**	**0.467 (0.180)**	**0.428 (0.170)**	**0.427 (0.228)**	**0.510 (0.300)**	**0.541 (0.328)**

In conclusion, skew-sensitive ensembles are benefited by considering different levels of imbalance and complexities for training the pool of base classifiers. And adapting the fusion function to the most recent operational imbalance employing the proposed scheme allows to provide a higher level of performance, mainly in the capacity of the system to preserve a low fpr.

4.5.4.2 Individual-Specific Analysis

Following an individual-specific analysis, Table 4.5 shows the average fpr, tpr, *precision* and F_1 performance measures for two of individuals enrolled to the system. The performance of

the eight test blocks with different imbalance levels are included, following the same structure as the Table 4.4. The levels of imbalance for each block are shown in the first row of Table 4.7.

Table 4.5 Average performance measures for different individuals enrolled to the system, setting a target 1% fpr on test blocks at different t times. The standard error is detailed between parenthesis

Approach	Measure	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
Module for Individual 58									
Imbalanced training+BC	fpr	1.71% (0.027)	1.73% (0.023)	3.18% (0.040)	2.63% (0.031)	3.96% (0.041)	3.27% (0.035)	2.98% (0.038)	3.06% (0.032)
	tpr *recall*	32.24% (0.320)	33.48% (0.391)	46.23% (0.360)	31.94% (0.329)	56.97% (0.333)	38.89% (0.351)	36.14% (0.358)	36.10% (0.390)
	precision	56.55% (0.370)	48.65% (0.284)	33.62% (0.242)	29.30% (0.222)	27.14% (0.179)	23.99% (0.161)	24.46% (0.160)	43.11% (0.263)
	F_1	0.351 (0.291)	0.317 (0.295)	0.331 (0.225)	0.270 (0.218)	0.318 (0.163)	0.255 (0.176)	0.248 (0.206)	0.307 (0.242)
Proposed approach	fpr	1.71% (0.027)	0.16% (0.002)	0.36% (0.003)	0.39% (0.004)	1.15% (0.005)	0.74% (0.005)	0.57% (0.004)	1.05% (0.009)
	tpr *recall*	32.24% (0.320)	29.41% (0.353)	48.96% (0.344)	24.07% (0.193)	66.76% (0.243)	32.71% (0.237)	31.75% (0.249)	33.47% (0.350)
	precision	56.55% (0.370)	77.72% (0.311)	68.22% (0.278)	63.57% (0.241)	49.26% (0.178)	47.01% (0.186)	50.74% (0.290)	47.28% (0.361)
	F_1	0.351 (0.291)	0.348 (0.366)	0.530 (0.330)	0.329 (0.224)	0.556 (0.184)	0.362 (0.211)	0.370 (0.251)	0.365 (0.343)
Module for Individual 209									
Imbalanced training+BC	fpr	9.97% (0.066)	8.04% (0.056)	5.79% (0.043)	7.45% (0.052)	3.64% (0.027)	4.32% (0.031)	4.66% (0.035)	7.93% (0.072)
	tpr *recall*	83.54% (0.285)	81.21% (0.291)	78.82% (0.287)	78.42% (0.274)	90.01% (0.212)	94.53% (0.172)	92.37% (0.213)	91.18% (0.228)
	precision	36.71% (0.225)	31.30% (0.209)	25.34% (0.140)	27.34% (0.136)	33.52% (0.175)	35.10% (0.183)	37.73% (0.223)	44.58% (0.254)
	F_1	0.489 (0.233)	0.421 (0.211)	0.363 (0.171)	0.386 (0.162)	0.469 (0.195)	0.491 (0.196)	0.507 (0.230)	0.557 (0.241)
Proposed approach	fpr	9.97% (0.066)	4.19% (0.022)	2.57% (0.012)	4.00% (0.021)	1.77% (0.008)	2.15% (0.014)	1.65% (0.011)	3.12% (0.030)
	tpr *recall*	83.54% (0.285)	84.36% (0.273)	84.18% (0.216)	83.81% (0.295)	89.98% (0.216)	95.20% (0.154)	93.82% (0.203)	91.30% (0.195)
	precision	36.71% (0.225)	40.55% (0.183)	40.17% (0.143)	39.39% (0.173)	44.80% (0.143)	49.18% (0.168)	56.83% (0.157)	63.42% (0.241)
	F_1	0.489 (0.233)	0.536 (0.201)	0.534 (0.158)	0.529 (0.209)	0.585 (0.156)	0.628 (0.148)	0.694 (0.166)	0.725 (0.208)

According to the initial performance presented by the system for individual 58, it can be categorized as a *goat*-like individual (see Table 4.3). For this individual, the tpr is initially low ($tpr < 55\%$) and maintained at that level for all test blocks except for $t = 5$. And the initially low fpr level that is very close to the desired 1%, is also maintained low through the operation

of the system. This evidences that the performance for this *goat*-like individual increased after the adaptation to the operational imbalance, but it remains in the same Doddington category with a low tpr, and also presents a low fpr regardless of the adaptation. It can also be seen that adapting the fusion function to the operational imbalance can potentially increase the tpr of the system at certain imbalances, as it happened for $t = 3$ and $t = 5$.

Similarly, according to the initial performance shown for individual 209, it can be categorized as a *lamb*-like individual. Although a high level of positive detections is presented by this individual-specific ensemble, a high level of negative acceptances is also shown in Table 4.5. This high fpr is significantly reduced after the system is adapted to the operational imbalance, becoming more accurate to discard the non-target samples. On the other hand, the tpr for this module is initially high $tpr > 80\%$, and is maintained or increased when the system is adapted to the operational imbalance. This shows the effectiveness of the system to maintain or even increase the amount of correct positive detections when the operational imbalance is taken into account. It also confirms the difficulty faced by the BC algorithm in the estimation of detection thresholds with balanced validation data.

4.5.4.3 Approximation of Operational Imbalance

As the operational imbalance changes over time, the system produces an estimate of such imbalance. In the SSBC algorithm, the accuracy of this estimation directly depends on the levels of imbalance considered in the initial set of imbalances Λ. This problem is avoided by the HDx and HDy methods. A sensitivity analysis was performed by variating the amount of imbalance levels in Λ, employing the balanced case and three other imbalanced cases. The imbalance space was evenly sampled adding five imbalances at the time. In that manner, the first set is composed of the balanced set plus 5 different imbalances ($\Lambda^1 = \{1:1, 1:20, 1:40, 1:60, 1:80, 1:100\}$), the second is composed of balanced and 10 imbalances ($\Lambda^2 = \{1:1, 1:10, 1:20, ..., 1:100\}$), the third is composed of balanced and 20 imbalances ($\Lambda^{20} = \{1:1, 1:2, 1:3, ..., 1:20\}$), and the last set contains 50 imbalances ($\Lambda^{50} = \{1:1, 1:2, 1:3, ..., 1:50\}$).

Table 4.6 Average performance measures for different sizes of Λ, for a desired 1% fpr on a test block with the maximum imbalance $\lambda^{max} = 1:100$. The standard error is detailed between parenthesis

Measure	Λ^1	Λ^2	Λ^{20}	Λ^{50}
fpr	1.59 % (0.18)	1.52 % (0.16)	1.53 % (0.15)	1.49 % (0.15)
tpr	58.41 % (5.62)	58.33 % (5.51)	58.47 % (5.32)	57.93 % (5.33)
precision	37.49 % (2.88)	38.46 % (2.78)	38.63 % (2.82)	38.73 % (2.94)
F_1	0.42 (0.04)	0.43 (0.03)	0.43 (0.03)	0.43 (0.03)

Table 4.6 presents the performance evaluated for the whole system using different resolutions of Λ, using a single test set with the maximum level of imbalance used in the experiment, $\lambda^{max} = 1:100$. Zooming to a more general scope, we can conclude that it may not be necessary to use the highest available resolution in terms of known levels of class imbalance (Λ), in order to obtain a good estimation of the operational imbalance.

The accuracy of the method based on the Hellinger distance employing the set Λ of validation sets was compared for different sizes of Λ (levels of imbalance employed by SSBC). The HDx and HDy quantification methods are included in the comparison in the scenario with real data. Table 4.7 presents the average imbalance estimated with three different sizes of Λ, and the HDx quantification method. The test includes the blocks of operational data for t=2, 5, 7 and 8 (from Table 4.4). Blocks for t=2, 8 were selected for its relatively small imbalance (1:35 or less), block for t=7 presents medium imbalance (1:50), and the block for t=5 presents the maximum imbalance employed in the experiments (1:100).

Results in Table 4.7 show increasing the size of Λ (adding imbalance levels) allows to increase the accuracy of the imbalance estimation. However there is a limit imposed by the characteristics of the histogram representations of the joint conditional probability of target and non-target samples. As it was seen in the synthetic experiment, as the probability of error

Table 4.7 Actual imbalance in test and the average number of ROIs for target individuals, as well as average imbalance estimated with the different lambda values and the HDx method (2 estimations per block - every 15 minutes)

t=2		**t=5**		**t=7**		**t=8**	
Imbalance in test blocks							
1:35		1:100		1:50		1:15	
Average target ROIs per block for 10 individuals							
65.50		116.40		113.30		95.00	
(4.62)		(5.87)		(5.84)		(6.44)	
Estimated imbalance, |Lambda| = 5							
1:59.6	1:69.7	1:59.8	1:79.5	1:74.3	1:61.0	1:65.9	1:60.8
(0.25)	(0.31)	(0.25)	(0.31)	(0.37)	(0.29)	(0.32)	(0.31)
Estimated imbalance, |Lambda| = 20							
1:51.0	1:61.8	1:55.8	1:72.8	1:69.1	1:53.1	1:60.8	1:54.1
(0.31)	(0.27)	(0.27)	(0.31)	(0.29)	(0.33)	(0.37)	(0.25)
Estimated imbalance, |Lambda| = 50							
1:47.3	1:57.7	1:53.2	1:68.9	1:64.9	1:51.0	1:55.5	1:51.9
(0.32)	(0.30)	(0.28)	(0.26)	(0.29)	(0.33)	(0.32)	(0.25)
Estimated imbalance, HDx							
1:9.1	1:9.8	1:8.5	1:10.5	1:13.4	1:11.7	1:12.0	1:11.3
(0.49)	(0.47)	(0.19)	(0.42)	(0.69)	(0.54)	(0.39)	(0.35)

increases (i.e. the uncertain overlapping zone between class distributions in the feature space), the joint conditional probability becomes more complex. Several non-target samples lying in the overlapping zone contribute to the histogram bins that correspond to the target class. Thus, the joint probability (histogram representation) of a data set with a high imbalance resembles the joint probability of a data set with lower imbalance. This phenomenon may be emphasized if a data management strategy is employed to select the most informative validation samples, like the one proposed in (De-la Torre *et al.*, 2013). That strategy is based on the KL divergence, and picks those target and non-target samples lying precisely in the region of overlap. However, if the data management strategy is reversed to include the less representative samples, the samples in the overlapping zone will be discarded. In this way, the samples that are useful for imbalance estimation can be separated from those that provide more information for class discrimination. However, this issue falls out of the reach of this chapter.

Comparing the approximation based on the Λ validation sets with the HDx quantification, the first shows a higher accuracy for imbalances close to 1:50, although fails for other cases. In the same sense, the HDx quantification is better for small imbalances (close to 1:15), although it fails to estimate greater operational imbalances. This phenomenon can be explained by the fact that HDx employs the whole set of validation samples, which provides the maximum imbalance but also the greatest amount of samples in the overlapping area. Anyhow, a more detailed analysis and comparison is required in order to evaluate which of the methods provide a better estimation in operations.

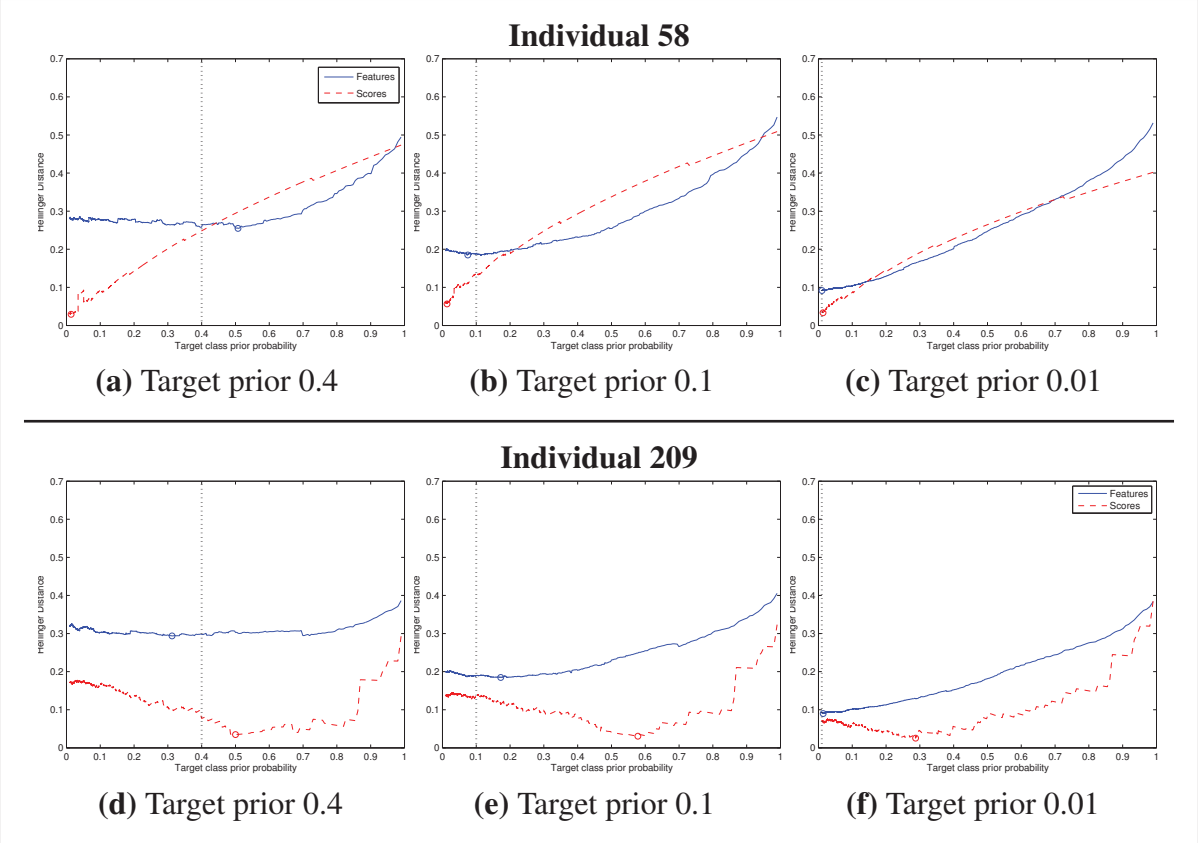


Figure 4.12 Hellinger distance between validation and test data from target and non-target distributions across different prior probabilities. The small circles correspond to the global minimum of the estimations, and constitute the approximation to the target prior probability. The experiment was realized with data from target individuals 58 and 209 and randomly selected non-target samples

A deeper characterization of the Hellinger distance between histogram representations of joint conditional probabilities for real samples is presented in Figure 4.12. The Hellinger distance was obtained by comparing a test set with different (but fixed) imbalances against randomly selected validation samples, and all the possible imbalances (prior probabilities) were covered. The curves shown in Figure 4.12 evidence the difficulty faced by both quantification methods to accurately estimate the imbalance of a set of test samples. Figures 4.12 a, b and c show that the imbalance for the goat-like individual 58 is more easily estimated employing the Hellinger distance in the feature space, and the score space produces less accurate estimations. However, both methods are accurate when the imbalance is high (target prior close to 0.01). This effect is related to the difficulty of the classification problem, as it was seen in Section 4.4 and discussed before. Figures 4.12 d, e and f show that the imbalance estimation for the lamb-like individual 209 is also challenging for both methods, that in all cases fail in finding the true target prior probability. This problem can be associated with the abundance of samples from wolf-like individuals, which lie precisely in the region that defines the target class in the feature space, and bias the imbalance estimation towards the target class. In any way, the Hellinger distance estimated in the feature space seems to provide a better estimation of the target prior probability.

Figure 4.13 shows the real and estimated imbalances for the same trajectory, with randomized ROIs for generalization purposes. The Λ sets employed in the simulation where Λ^1, Λ^4, Λ^{10}, with 5, 20 and 50 levels of imbalance respectively, and the HDx quantification method. The operational imbalance was estimated every 3 minutes with a window that considers operational data for the last 15 minutes of captures and corresponds to the black dashed line. The true imbalance estimated over time corresponds to the red solid line. It can be seen that the estimation of class imbalance for the first minutes falls to zero in the four cases, which is related to the initial state of the system with an empty buffer of operational samples. The highest peak in the curve for true class imbalance was chosen for a visual comparison, which appears close to minute 140. The blue ellipses in the four graphs show the estimated imbalance levels, showing that the best fit between real and estimated imbalances is given by the HDx quantification, with

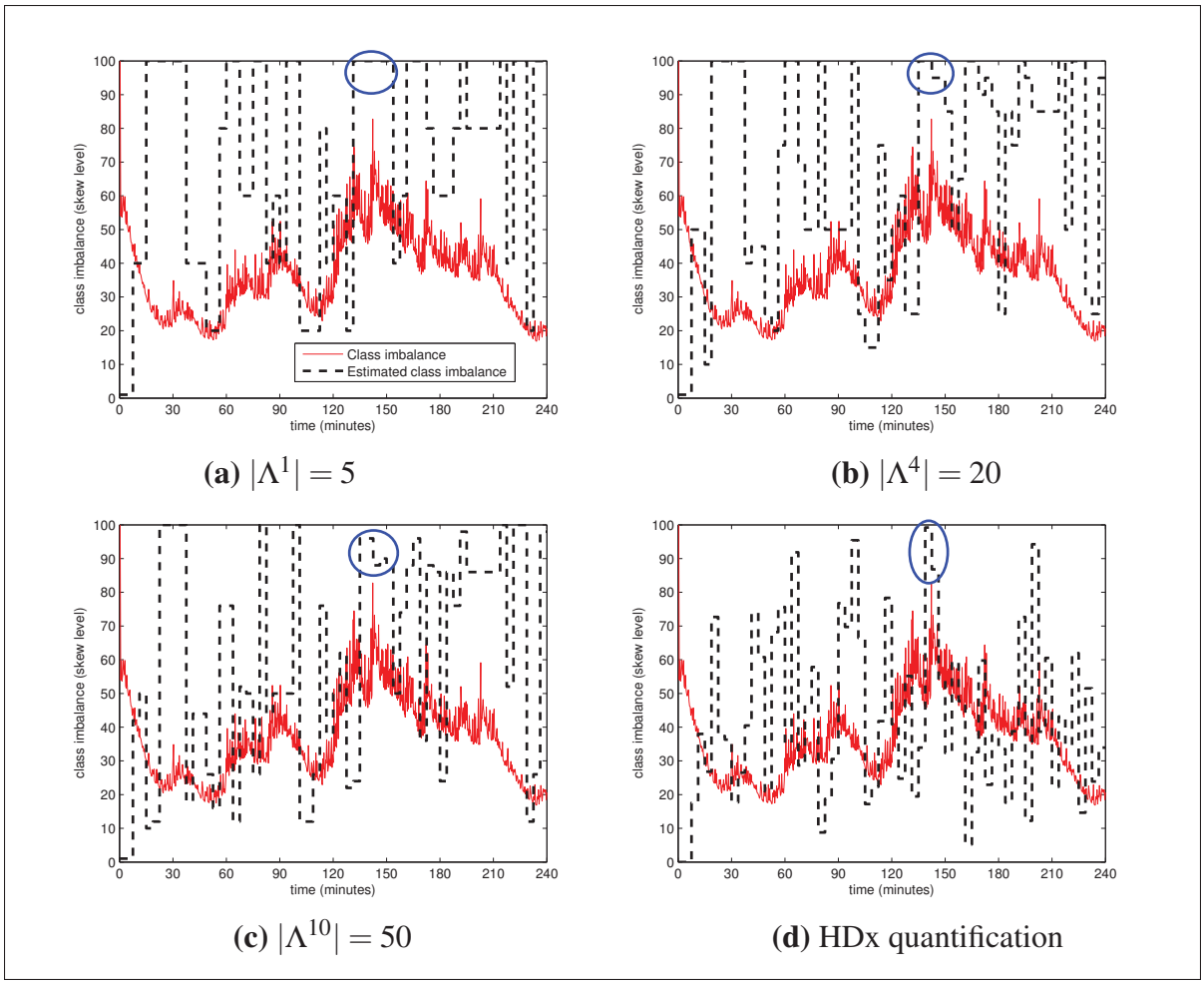


Figure 4.13 Adaptation of the level of class imbalance over time, corresponding to individual 58 at the first experimental trial. Comparison of four different sizes of $|\Lambda|$ corresponding to 5 (a), 20 (b) and 50 (c) levels of imbalance, for an evenly sampled space of imbalances between 1:1 and 1:100

a narrower peak closer to the solid red graph. However, this tendency is not always true, as can be seen looking at the peak of the black dashed line that appears between minutes 90 and 100 minutes in the four cases, indicating that the estimated imbalance was better with any of the Λ sets. This shows that even though the HDx quantification performs better than the raw comparison of Hellinger distance between operational and validation histograms, there is a limit in the estimation related to the data used in validation. In any case, the superiority of the HDx quantification is evidenced by the more objective comparison shown below.

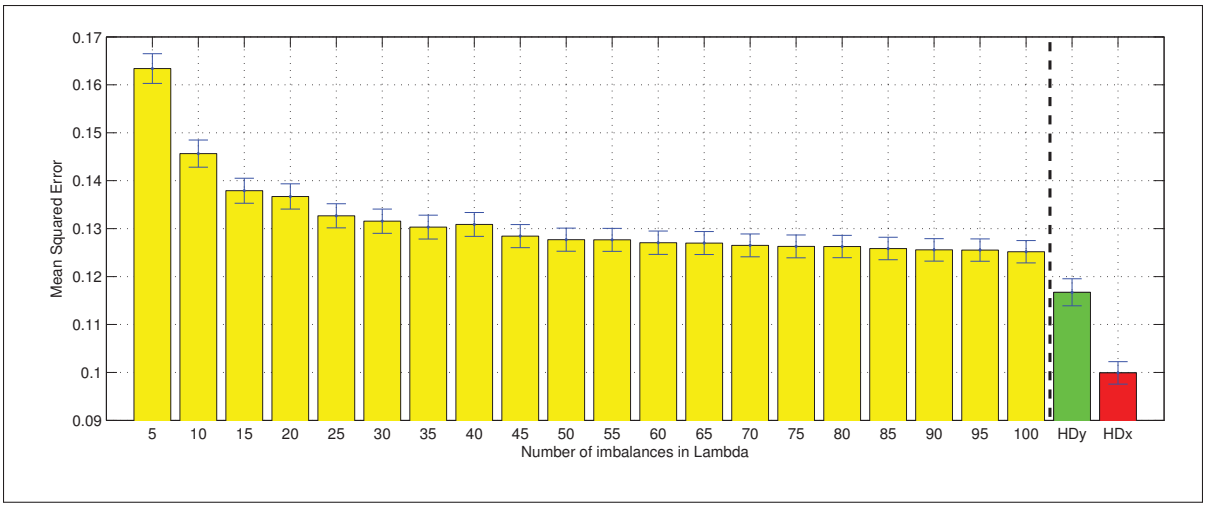


Figure 4.14 Average mean squared error between real and estimated operational imbalances for different number of imbalance levels in Λ for the method based on different validation sets, compared to the HDx and HDy quantification (right extreme)

A numeric estimation of the difference between the real and estimated imbalance curves is the mean squared error (MSE), which is widely employed in statistics to measure the average of the squares of the differences between the estimation and the quantity that is estimated. Figure 4.14 presents the average of the mean squared error between the true and estimated imbalances for all the twenty different resolutions used in the experiment using the method that employs different validation sets (Radtke *et al.*, 2013b), and the HDy and HDx quantification methods (Gonzalez-Castro *et al.*, 2013). The results involve the 24 replications of the experiment and the 10 individuals of interest. In the Figure 4.14, the mean of the MSE observed in the first bar that corresponds to 5 levels of imbalance in Λ is close to 0.162, and drops to 0.137 when 20 levels of imbalance are employed. After using 20 levels of imbalance, the reduction in the MSE for more levels of imbalance in Λ is not significant but consistent, as evidenced by a median of 0.128 and 0.125 for 50 and 100 imbalances respectively. Finally, the HDy and HDx quantification methods present a significantly lower average MSE of 0.117 and 0.101 respectively.

A Kruskal-Wallis analysis on the complete set of results using the original approximation method (first 20 boxplots) throws a p-value of $6.82 \times 10^{-29} \leq 0.05$, which confirms that the

differences in MSE between the estimated and real imbalances are significant with a 95% confidence interval. The same analysis on the last 17 test cases using the original method (removing Λ^1 with 5, 10 and 15 levels of imbalance) throws a higher p-value of $0.1193 > 0.05$, which means that there's no significant difference between all the last 17 cases with a confidence interval of 95%. However, pairwise Kruskal-Wallis analysis for $(\Lambda^4, \Lambda^{20})$ and $(\Lambda^5, \Lambda^{20})$ produce p-values of $0.0021 < 0.05$ and $0.0436 \not< 0.05$, confirming a significant difference. Thus, according to these results using more levels of imbalance in Λ provides significantly higher resolution for imbalance estimation. Finally, the Kruskal-Wallis test between the original imbalance estimation method with Λ^{20} and the HDx quantification throws a p-value of $3.94 \times 10^{-17} \ll 0.05$, showing a significant superiority of the HDx quantification method when compared the method based on several validation sets.

4.5.5 Trajectory-Level Analysis

In this scenario videos were concatenated one after the other, emulating a passport checking scenario where individuals approximate to the camera one after the other from the waiting line. Four blocks of 30 minutes were obtained (D_1, D_2, D_3 and D_4), showing different imbalances in a realistic scenario. The first two blocks are composed of trajectories from capture session 2, and the last two blocks are composed of trajectories from capture session 3. Trajectories from blocks D_1 and D_3 were captured with an unzoomed camera, and trajectories from blocks D_2 and D_4 were captured with a zoomed camera. The four blocks were presented to the system in order.

Table 4.8 shows the average performance of the system using balanced BC and SSBC for the passport checking scenario, after selecting the operations point at $fpr = 1\%$. It can be seen that the performance of the proposed approach is significantly higher than the performance for the reference system. And comparing the fpr for both systems, it can be seen that the performance superiority of the proposed approach is mainly due to its capacity to keep a low amount of false alarms after the operations point is adapted to the operational imbalance. This capacity proposed approach is related to the employment of the widely available non-target samples

Table 4.8 Average performance measures for different approaches for an $fpr = 1\%$ on test blocks at different t times. The standard error is detailed between parenthesis, and bold numbers symbolize significant difference in terms of F_1 measure with respect to the reference system

Approach	**Measure**	$t = 1$	$t = 2$	$t = 3$	$t = 4$
Reference system	fpr	5.15% (0.025)	4.15% (0.024)	4.71% (0.023)	3.30% (0.014)
	tpr *recall*	61.54% (0.171)	56.94% (0.234)	59.74% (0.283)	59.41% (0.313)
	precision	23.19% (0.077)	24.67% (0.099)	30.61% (0.154)	34.43% (0.171)
	F_1	**0.300 (0.094)**	0.307 (0.135)	0.363 (0.183)	0.383 (0.217)
Proposed approach	fpr	5.15% (0.025)	1.47% (0.010)	1.61% (0.013)	1.11% (0.006)
	tpr *recall*	61.54% (0.171)	54.60% (0.327)	49.79% (0.341)	54.40% (0.354)
	precision	23.19% (0.077)	40.82% (0.158)	48.82% (0.251)	48.13% (0.247)
	F_1	**0.300 (0.094)**	**0.422 (0.204)**	**0.434 (0.238)**	**0.477 (0.285)**

to establish the decision frontier at the combination function, enhancing the discrimination between target and non-target classes.

The face trajectories built using the IVT face tracker to regroup target facial regions were used for trajectory-based analysis of the system in this real passport-checking scenario. The first time a face is found in the video sequence, the location of the facial region is employed to initialize the tracker that follows it until the individual leaves the scene. Target predictions produced by the system were accumulated over time for full trajectories to provide overall decisions, and the detection threshold was applied to these accumulations.

Figure 4.15 presents an example of the accumulation of detections produced by the EoD trained on samples from individual 151, for the sequence of individuals entered in the scene over time. Two zoomed regions that are representative of the system response are also shown in the same figure. The accumulation of positive predictions produced in response to the target

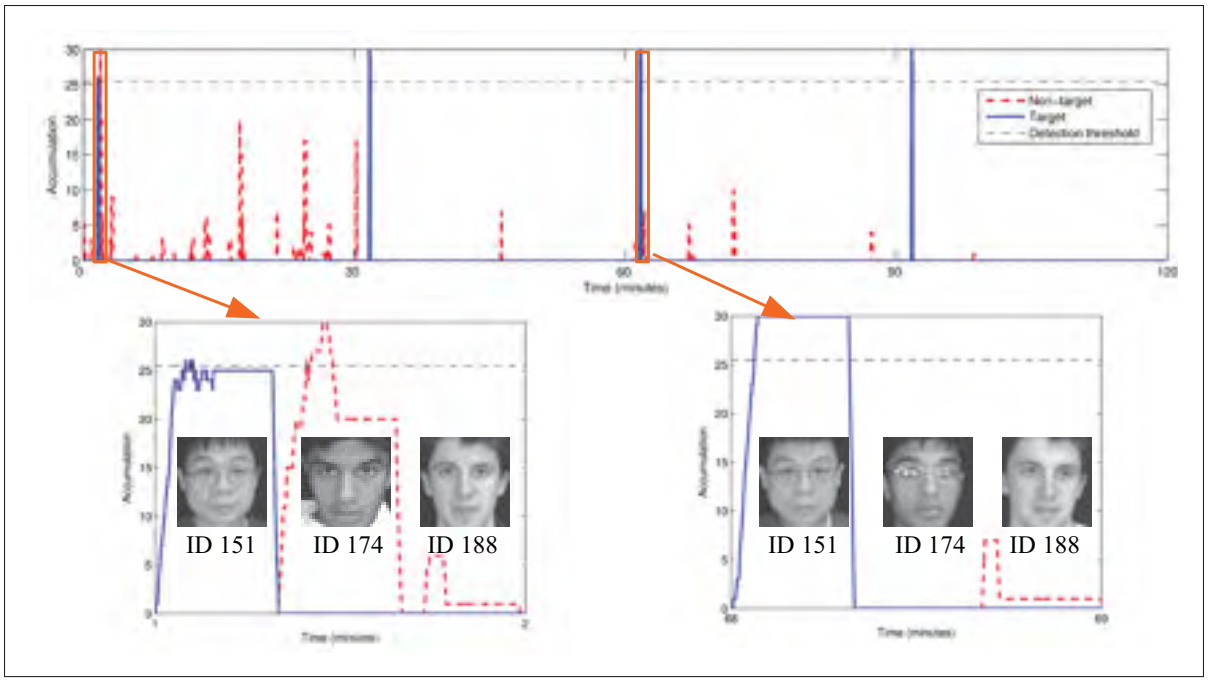


Figure 4.15 Examples of target detection accumulations for concatenated input trajectories corresponding to the module trained for individual 151. The left and right zoomed views of the graph show the target individual entering in the scene, as well as two non-target individuals with ID 174 and 188

trajectory are drawn in a bold, solid blue line, and the accumulations for non-target trajectories are drawn in bold, dashed red line. The detection threshold is drawn with a dashed black horizontal line. Target and non-target trajectories produce accumulation levels that may surpass the detection thresholds, producing true and false positive detections. In the left zoomed area in Fig. 4.15, the target trajectory was correctly detected, whereas one of the non-target trajectories was incorrectly recognized as belonging to the target individual. In the right zoomed area in Fig. 4.15, the target trajectory was detected with a higher accumulation than the initial left zoomed area, and the non-target trajectories were correctly rejected showing an increased discrimination after adapting the system to the operational imbalance.

As followed in the protocol, the first adaptation of the fusion function is performed after 30 minutes of operation, where the last block of operational samples are used for imbalance estimation. When the first capture session is presented, the discrimination between target and

non-target trajectories is less clear, evidenced by some false positive detections (see the left zoomed area in Fig. 4.15). When the operation point is adapted –after minute 30–, the system increases its capacity to discriminate between target and non-target trajectories, as shown in the right zoomed area in Fig. 4.15. This is a clear evidence that selecting the operations point based on a validation set with the appropriate class imbalance allows for a better discrimination between target and non-target classes, which is extended to the overall trajectory-based response of the system.

Table 4.9 Average operational imbalance and overall AUC-5% for the reference system and the proposed approach, considering the 10 individuals over 24 trials. The standard error is shown in parenthesis

	t=1	**t=2**	**t=3**	**t=4**
Average Imbalance	1:15.73	1:16.02	1:10.14	1:15.16
Average target ROIs	85.3 (7.07)	102.7 (6.56)	79.3 (5.35)	95.0 (6.44)
Reference system (AUC-5%)	67.87 (2.21)	67.67 (2.40)	71.41 (2.36)	73.36 (2.28)
Proposed approach (AUC-5%)	67.87 (2.21)	79.45 (1.98)	78.61 (2.14)	74.07 (2.57)

Table 4.9 shows the average operational imbalance, as well as the average overall AUC for the ROC curves obtained over $0 \leq fpr \leq 0.05$ (AUC-5%). The performance of the system for the first test block, when the operational imbalance is not considered, is significantly lower in terms of AUC-5%, compared to the performance after adapting the fusion function. This is the same tendency as the observed in the transaction-based evaluation, which confirms that the performance increase of the system using proposed approach can be extended to the overall system performance in video-to-video FR.

4.6 Conclusion

In video surveillance, it is often assumed that the proportions of faces captured for target and non-target individuals are balanced, known a priori and do not change over time. Recently,

some techniques have been proposed to adapt the fusion function of an ensemble according to class imbalance measured on operational data. However, skew sensitive ensembles commonly employ balanced training data to generate diverse pools of base classifiers, limiting the potential diversity produced using the abundant non-target data, with multiple levels of imbalance and complexity.

In this chapter, skew-sensitive adaptive classifier ensembles have been investigated and applied to video-to-video FR in video surveillance applications. The proposed scheme allows to combine classifiers trained by selecting data with varying levels of imbalance and complexity, and leads to a significant improvement of system's accuracy and robustness. During enrollment, target facial captures from a reference trajectory are combined with selected captures from non-target trajectories to generate a pool of 2-class classifiers using data with various levels of imbalance and complexity. During operations, face captures of each person in the scene are tracked and regrouped into trajectories for video-to-video FR, producing enhanced discrimination between target and non-target trajectories. The level of imbalance is periodically estimated from the input data stream using the HDx quantification, and pre-computed histogram representations of imbalanced data distributions. Finally, pre-computed histograms and ensemble fusion functions are updated based on the imbalance and complexity of operational data.

Results on synthetic problems show that the combination of the classifiers trained with different imbalance levels and complexities increases ensemble diversity and robustness, leading to an increase in the ROC and precision-recall performances. A comparison of imbalance quantification based on Hellinger distance in score and feature spaces shows that feature-based estimation is more accurate when the probability of error is high. Similarly, results on the CMU-FIA video data show that the proposed method can outperform other techniques in imbalanced environments. In that sense, transaction-based analysis shows a significantly higher performance in terms of F_1 measure, that is consistently higher for different operational imbalances. Individual-specific analysis indicates that goat- and lamb-like individuals can benefit the most from adaptation to the operational imbalance. Trajectory-based analysis shows that the

improvement presented at transaction level is propagated to the overall performance evaluated in a realistic video-to-video FR scenario.

The future work should consider exploiting the class imbalance at decision fusion level, setting imbalance-specific thresholds for the estimated test skew. Although HDx quantification method provided the highest accuracy with respect to the compared methods, there is still room for further improvement. Further characterization of the system in different and more challenging scenarios would be interesting, including for instance crowded and outdoor places. Other applications like gait-based biometrics may also be benefited from the findings of this research, since several individuals appear in videos. Finally, adaptation to permanent changes in the probability distribution of data due to changes in facial appearance may be addressed employing self-update techniques, leading to further improvement in the performance of the system.

GENERAL CONCLUSION

Systems for face recognition (FR) in video surveillance are applied in a range of scenarios like watchlist screening, face re-identification and search and retrieval. Several challenges are present in these applications, including the common assumption that the facial appearance of target individuals do not change over time, and that the proportions of faces captured for target and non-target individuals are balanced, known *a priori* and remain fixed. However, faces captured during operations vary due to capture conditions, the proportions of target and non-target individuals continuously change during operations, and facial models used matching are commonly not representative since they are designed a priori, with a limited amount of reference samples that are collected and labeled at a high cost.

In this Thesis, a framework for adaptive systems for video-to-video face recognition (FR) in video surveillance is proposed, contributing with new techniques to adapt the facial models for enrolled individuals of interest. This framework allows the systems for trajectory-based self-updating to automatically update facial models, considering gradual and abrupt changes in the classification environment. Besides, with the use of a modification to SSBC, the systems are capable to adapt the individual-specific ensembles to the operational imbalance.

In **Chapter 1**, a review on the most recent advances in adaptive video-to-video FR for video surveillance is described. It was found that adaptive multiple classifier systems (MCSs) have been successfully applied to video-to-video FR, where ensembles of 2-class Fuzzy ARTMAP classifiers, employing a DPSO strategy to generate a pool of classifiers with optimized hyperparameters, and Boolean combination (BC) to merge their responses in the ROC space. Besides, active skew-sensitive ensembles were recently proposed to adapt the fusion function according to the class imbalance measured on operational data. Finally, face tracking can be used to regroup the system responses linked to a facial trajectory (facial captures from a single person in the scene) for robust spatio-temporal recognition, and to update facial models over time using operational data.

In **Chapter 2**, the baseline framework is described. In this framework, the face of each target individual is modeled using an ensemble of 2-class classifiers, and integrates information from a face tracker and individual-specific ensembles for robust spatio-temporal recognition and for efficient self-update of facial models. Facial models are updated with all target samples extracted from highly confident trajectories (facial captures from a single person in the scene) are combined with non-target samples selected from the cohort and universal models. A learn-and-combine strategy is employed to avoid knowledge corruption and a memory management strategy based on Kullback-Leibler divergence is used to rank and select the most relevant target and non-target reference ROI samples for validation. Proof of concept validation has been performed on the CMU-FIA video dataset. Results show the response of proposed systems to gradual changes in facial appearance of individuals, as found in video surveillance, under semi-controlled or uncontrolled capture conditions. Transaction-level analysis shows that the proposed approach outperforms baseline systems that do not adapt to new trajectories, and provides comparable performance to ideal systems that adapt to all relevant target trajectories, through supervised learning. Subject-level analysis reveals the existence of individuals for which self-updating ensembles with unlabeled facial trajectories provides a considerable benefit. Trajectory-level analysis indicates that the proposed system allows for robust spatio-temporal video-to-video FR, and may therefore enhance security and situation analysis in video surveillance.

In **Chapter 3**, a particular implementation of the system has been characterized in a scenario with gradual and abrupt changes in the probability distribution of faces in feature space. This implementation consists in a pool of Probabilistic Fuzzy ARTMAP classifiers generated using a DPSO learning strategy. The classifiers are trained using the target samples from reference trajectories, and a set of non-target samples selected from the cohort and universal models using One-Sided Selection. The individual-specific pools of classifiers are combined using Boolean combination. Each ensemble seeks to recognize target individuals and self-update facial models based on facial trajectories defined by the tracker, tunning up individual-specific parameters for classification and decision fusion. Transaction-level results show that the proposed system

allows to increase AUC accuracy by about 3% in scenarios with abrupt changes, and by about 5% in scenarios with gradual changes. Subject-based analysis reveals the difficulties of face recognition with different poses, affecting more significantly the lamb- and goat-like individuals. Compared to reference spatio-temporal fusion approaches, results show that the proposed accumulation scheme produces the highest discrimination. The characterization of the system under abrupt (pose) and gradual (aging) patterns of changes indicate that the proposed system allows for improved overall transaction-level performance after self-update with operational face trajectories. Subject-level analysis reveals the difficulties faced to recognize the individuals under different face poses, affecting most significantly the performance of lamb- and goat-like individuals. A comparison between different spatio-temporal fusion approaches shows that the proposed scheme produces higher trajectory-based pAUC (5%) accuracy than other approaches, even for different window sizes. An analysis of the updates achieved by the system shows that by virtue of the increased discrimination, it presented a low number of incorrect updates even with the large number of non-target trajectories presented to the system during simulations.

In **Chapter 4**, skew sensitive adaptive ensembles of classifiers were investigated and applied to video-to-video face recognition in video surveillance. In the proposed scheme, classifiers are trained by selecting data with different levels of imbalance and complexities, leading to a significant improvement of the system's robustness and performance. During operations, face captures of an individual are tracked and regrouped to form face trajectories, employed for spatio-temporal recognition. The level of operational imbalance is periodically estimated from input data stream using the HDx quantification, and the fusion function as well as the pre-computed histogram representations of imbalanced data distributions are updated. Results on synthetic problems show that the combination of the classifiers trained with different levels of imbalance and complexity allows to increase ensemble diversity, and ROC and precision-recall accuracy. Subject-based analysis shows that goat- and lamb-like individuals are greatly benefited from adaptation to the operational imbalance. Finally, the system was successfully applied for skew-sensitive video-to-video FR.

Future Work

Although the proposed system demonstrated efficient adaptation in changing environments, it is very complex and more strategies may be required to control resources growth. In that sense, an important pending issue is to assess the scalability of the proposed system when the number of target individuals grows, as well as the number of cameras that capture the scene. In this case, resources were not limited, but pruning techniques may be employed to remove not relevant classifiers. In practice, the system should exploit internal knowledge (age, performance relevance, etc.) to remove some older or redundant classifiers over time. Moreover, the exploitation of the diversity of opinions should be guided by intelligent strategies, that validate the amount of classifiers used in the ensemble, and which of them are more useful according to a trade off between resources and accuracy. Change detection strategies may be employed to limit the number of classifiers added when self-update is activated.

Up to now, the system was characterized in environments with gradual and abrupt changes, but it would be interesting to analyze the performance of the system in an environment where multiple individuals are simultaneously present in scene. The use of skew-sensitive ensembles in video-to-video face recognition has shown to reduce the number of false positives, and combining these ensembles with self-update techniques may be a potential tool to reduce the number of false updates. Although HDx quantification provided the highest accuracy with respect to the compared methods, there is still room for improvement, and other techniques may be explored for the estimation of operational imbalance. Finally, other applications like gait or keystroke dynamics may also benefit from the findings of this research, which are biometric characteristics that also suffer of changes according to the age or certain health conditions.

APPENDIX I

SYNTHETIC EXPERIMENT ON RELEVANCE MEASURES

Two synthetic 2-class problems were designed to characterize the relevance measures in the 1D space. Fig. I-1 shows the original probability distributions used to generate the data for experiments. The central Gaussian distribution in both problems generates the positive samples, with a center of mass $\mu_2 = 0.5$. The centers of mass of the negative Gaussian distributions in Fig. I-1 (a) are $\mu_1 = 0.2$ and $\mu_3 = 0.8$, and in Fig. I-1 (b) the negative samples are randomly drawn from the 1D space according to a uniform distribution. All Gaussian distributions are characterized by a fixed variance of $\sigma = 0.01$. An ensemble of 7 PFAM classifiers has been trained for both problems on a balanced training set. A learning strategy based on DPSO is used for generation of base classifiers and co-jointly optimize all PFAM parameters, as proposed in (Connolly *et al.*, 2012). Classifier fusion is performed using BC.

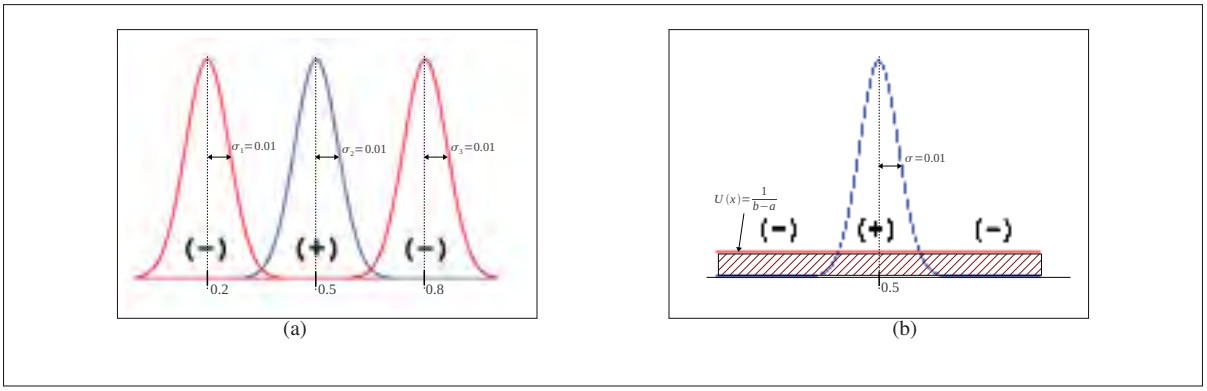


Figure-A I-1 Data distributions used to generate the training data for problems 1 (a) and 2 (b). In both figures the Gaussian distribution at the center generates the positive (+) samples, and the left and right distributions generate the negative (-) samples

The value of relevance measures for the PFAM ensembles corresponding to both problems are presented in Fig. I-2. Whereas the extension of surprise (average surprise) follows a shape similar to that of the surprise estimated for a single model, other measures focus on the overlapping of data distribution zones. Vote entropy uses decision level information (level D from

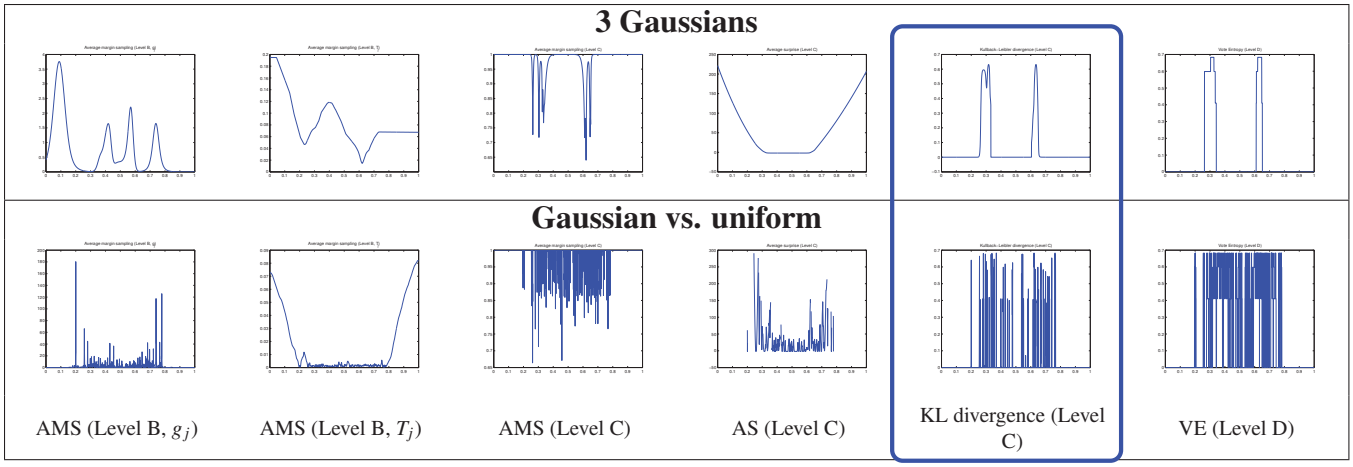


Figure-A I-2 Value of relevance measures obtained over the input space with an ensemble of 2-class PFAM classifiers for the *3 Gaussians* (top) and *Gaussian vs. uniform* (bottom) problems. From left to right, average margin sampling (AMS) at level B on g_j, AMS at level B on T_j, AMS at score level, average surprise (AS) at score level, Kullback-Leibler (KL) divergence at score level, and vote entropy (VE) at prediction level

Fig. 2.2), and hence presents a lower resolution (e.g. fewer ranking values). While KL divergence and average margin sampling both present a good resolution, the smoothness of curves for KL divergence, provide a better representation of the overlapping area.

APPENDIX II

FULL UPDATE TABLE IN A PROGRESSIVE TEST-UPDATE SCENARIO

Table 2.1 presents the details of the updates for the 10 independent replications of the experiment for the individuals of interest enrolled to the system, with the EoD_{ss} (PFAM) $\mathrm{LTM}_{KL,\lambda_k=100}$.

Table 2.1 IDs corresponding to the trajectories that surpassed the update threshold and were used for updating the selected EoDs on different replications (r) of the experiment (EoD_{ss}, $LTM_{KL,\lambda_k=100}$). Bold numbers correspond to trajectories selected for correct updates, and conflicts are marked with a box around the ID of the trajectory

Replic.	EoD_{ss} 2	Mod. 58	Mod. 72	Mod. 92	Mod. 147	Mod. 151	Mod. 176	Mod. 188	Mod. 190	Mod. 209
Update trajectories in D_1										
r=1	**2**	**58**	**72**	-	**147**	-	-	6,60,186,**188**,193,224	-	**209**
r=2	**2**	**58**	**72**,179	**92**	-	-	-	**188**,224	**190**	**209**
r=3	**2**	**58**	**72**	**92**,235	**147**	**151**	-	**188**	136,**190**	**209**
r=4	**2**	**58**	**72**	**92**,235	**147**	-	-	**188**	-	**209**
r=5	**2**	**58**	**72**,179	**92**	**147**	-	-	**188**,224	-	**209**
r=6	**2**	**58**	**72**	**92**	**147**	**151**	-	**188**	**190**	**209**
r=7	-	**58**	**72**,179	-	**147**	-	-	**188**,224	**190**	**209**
r=8	**2**	**58**	**72**,179	**92**	**147**	**151**	**176**	**188**	-	**209**
r=9	**2**	**58**	**72**	-	**147**,222	**151**	**176**	**188**,224	**190**	**209**
r=10	**2**	**58**	**72**,179	-	**147**	**151**	**176**	**188**	-	**209**
Update trajectories in D_2										
r=1	220	-	136,175	-	**147**	-	-	**188**	136	**209**
r=2	220	-	179	-	-	-	-	104,**188**	99,127,136,**190**,201	-
r=3	-	**58**	-	-	**147**	-	-	-	127,136,**190**	**209**
r=4	-	-	148	-	**147**	-	-	-	136	**209**
r=5	-	**58**,134	23,148,175	-	**147**	-	-	**188**	136	**209**
r=6	220	**58**	-	-	**147**	**151**	**176**	104,**188**	136,**190**	**209**
r=7	-	-	-	-	**147**	-	**176**	**188**	136,**190**,197	**209**
r=8	-	**58**	-	-	**147**	-	**176**	104,122,**188**	134,136	**209**
r=9	-	-	134	-	**147**	**151**	**176**	104,151,153,**188**	99,136,**190**	**209**
r=10	-	**58**	94	-	**147**	**151**,174	-	104,**188**	136	**209**
Update trajectories in D_3										
r=1	**2**	-	136	37,**92**,134,148	-	140,**151**	3	**188**	148,**190**	**209**
r=2	**2**,108	-	179	**92**,134,148	-	**151**	140,151	**188**	136,**190**	**209**
r=3	**2**	**58**	179	134,148	-	92,107,**151**,202	-	**188**	136,148,**190**	**209**
r=4	**2**	-	179	**92**,134,148	-	**151**	108,151,177	-	47,84,136,**190**	**209**
r=5	**2**	**58**	-	134,148	-	**151**	151	**188**	47,84,136,148,**190**	**209**
r=6	-	**58**	37,179	37,**92**,134,148	-	**151**	**176**,177	**188**	12,47,58,84,136,148,**190**	**209**
r=7	**2**	-	179	37,**92**,134,148	-	**151**	-	**188**	136,148,**190**	**209**
r=8	**2**	-	37,179	**92**	-	107,**151**	**176**	**188**	84,136,**190**	**209**
r=9	**2**	-	84,134,148	58,**92**	-	**151**	-	**188**	136,**190**	**209**
r=10	**2**	**58**	37,179,197	148	-	11,**151**	140	-	84,136,**190**	-

APPENDIX III

FULL UPDATE TABLES IN A SCENARIO WITH GRADUAL AND ABRUPT CHANGES

Tables 3.1 and 3.2 present the correct and incorrect update trajectories used by the system for self-update. The tables correspond to the scenarios with abrupt (pose) and gradual (aging) changes respectively.

Table 3.1 Update table for the system with correct (bold) and incorrect update trajectories in the Left and Right update trajectories

Update trajectories in D_R

Replication	EoD_{ss} (2)	EoD_{ss} (3)	EoD_{ss} (21)	EoD_{ss} (58)	EoD_{ss} (72)
r = 1	-	-	-	11,13,21,22,34,37,43,48,**58**, 64,74,79,82,99,121,130,155, 162,188,193,195,201	11
r = 2	-	-	**21**,188	11,48,**58**,74,82,99	7,11,22,37,58,62, 79,124,136,190,201,213
r = 3	-	-	**21**,34,48,82,99,162,188	11,13,21,22,43,48,**58**,74,79, 82,99,162,188,195,201	-
r = 4	-	-	16,206	11,13,16,22,34,37,48,**58**,72, 74,79,80,82,92,157,188,206	-
r = 5	-	-	16,**21**,64	11,48,82	-
r = 6	-	-	-	-	7,213
r = 7	-	-	**21**,34,37,48,58,82,99, 162,188	2,7,11,21,22,34,37,43,48,**58**,64, 74,79,82,99,121,130,156,157,162, 167,188,190,193,195,201	
r = 8	-	-	**21**,188	11,21,22,34,37,48,**58**,74,79,82,99, 162,188,195,201	-
r = 9	-	-	**21**,34,48,82,162,188	11,22,48,**58**,74,82,99	-
r = 10	-	-	**21**,34,37,48,58,64,74,99, 130,162,188	11,22,48,**58**,74,82,99,162	-
Replication	EoD_{ss} (99)	EoD_{ss} (121)	EoD_{ss} (188)	EoD_{ss} (190)	EoD_{ss} (213)
r = 1	-	-	-	-	-
r = 2	-	-	-	-	-
r = 3	-	-	-	16	-
r = 4	-	74,**121**	-	-	-
r = 5	-	-	**188**	-	-
r = 6	-	-	-	-	-
r = 7	-	-	60,67,165,209	-	-
r = 8	-	-	2,3,7,34,60,62,67,73,107, 121,133,167,**188**,201,202,209	-	-
r = 9	-	-	-	-	-
r = 10	-	-	-	-	-

Update trajectories in D_L

Replication	EoD_{ss} (2)	EoD_{ss} (3)	EoD_{ss} (21)	EoD_{ss} (58)	EoD_{ss} (72)
r = 1	-	-	-	176	-
r = 2	-	-	-	-	-
r = 3	-	-	-	-	-
r = 4	-	-	-	-	-
r = 5	-	-	-	-	-
r = 6	-	-	-	-	130
r = 7	-	-	-	-	-
r = 8	-	-	99,162,209	124	-
r = 9	-	-	-	3,121	-
r = 10	-	-	-	-	-
Replication	EoD_{ss} (99)	EoD_{ss} (121)	EoD_{ss} (188)	EoD_{ss} (190)	EoD_{ss} (213)
r = 1	-	-	-	-	-
r = 2	-	-	-	-	-
r = 3	-	-	-	-	-
r = 4	-	-	-	-	-
r = 5	-	-	-	-	-
r = 6	-	-	-	-	-
r = 7	-	-	-	-	-
r = 8	-	-	3,130	-	-
r = 9	-	-	-	-	-
r = 10	-	-	-	-	-

Table 3.2 Update table for the system with correct (bold) and incorrect update trajectories in the Left and Right update trajectories

Update trajectories in D_2

Replication	EoD_{ss} (2)	EoD_{ss} (3)	EoD_{ss} (21)	EoD_{ss} (58)	EoD_{ss} (72)
r=1	**2**,41,220	**3**,43,74	**21**,31,34,56,118	**58**,92,134,148,235	23,**72**,94,99,127, 175,197,201,206,209
r=2	**2**,220	**3**,43,74	118	37,**58**,92,102,134, 148,235	23,**72**,127,201
r=3	**2**,91,220	**3**,13,43,74	**21**,31,49,118,132	**58**,92,148,235	23,**72**,99,127,148, 201,206,209
r=4	**2**,91,220	**3**	**21**,31,56,70,71, 132,207,227	37,**58**,92,102,134, 148,235	19,23,**72**,94,148, 175,198,201,206,209
r=5	**2**,220	**3**	**21**,31,118,207	**58**,92,134,148	**72**
r=6	220	**3**,43,74	**21**,31,34,56,70, 102,118,132,227	37,**58**,92,134,148, 235	23,197
r=7	220	**3**,74	**21**,31,34,118	**58**,92,134,148	**72**
r=8	**2**,220	**3**,43,74	**21**,34,207	37,**58**,92,134,148, 235	23,**72**,127
r=9	**2**,220	**3**	**21**,31	**58**	23,127,206
r=10	**2**,220	**3**	-	**58**,92,134,148,235	-
Replication	EoD_{ss} (99)	EoD_{ss} (121)	EoD_{ss} (188)	EoD_{ss} (190)	EoD_{ss} (213)
r=1	-	73,**121**	29,49,83,96,104, 122,140,147,179,**188**	127,136,157,175, **190**,197,201	**213**
r=2	**99**,106,136,190	73,**121**,170	96,104,122,**188**	58,127,136,157,**190**, 201	147,157,**213**
r=3	12,**99**,106,136,157, 175,190,201,229	**121**	96,104,122,179,**188**	127,136,**190**,197,201	188
r=4	-	73,**121**,123,170	104,122,**188**	80,127,136,157,**190**, 197,201	49,**213**
r=5	**99**,106,136,190	**121**,123	96,104,122,140,179, **188**	136,**190**	131,147,**213**
r=6	-	**121**	147,**188**	136,157,**190**,197,201	147,157,**213**
r=7	**99**,106,136,190	**121**,123	147,**188**	12,127,136,157,175, **190**,197,201	49,106,147,157,188,**213**
r=8	**99**,106,136,157, 175,190,197	57,73,108,114, **121**,123,170	104,122,**188**	136,**190**,197	188,**213**
r=9	**99**,106,136,190	73,**121**,123	-	127,136,**190**,201	**213**
r=10	**99**	**121**,170	96,104,122,**188**	99,127,136,**190**,201	**213**

Update trajectories in D_3

Replication	EoD_{ss} (2)	EoD_{ss} (3)	EoD_{ss} (21)	EoD_{ss} (58)	EoD_{ss} (72)
r=1	**2**	**3**	**21**	37,92,134,148	58,84,148,157
r=2	**2**,41	-	-	37,**58**,84,92,102, 134,148	2,43,108,118
r=3	**2**	**3**,154,186	**21**,49,91,118,202, 213	**58**,157	-
r=4	**2**	**3**	16,**21**,140	37,**58**,92,134,148, 157,206	198
r=5	**2**	**3**	**21**	37,**58**,84,92,102, 134,148,157	148
r=6	**2**,113	**3**,11,30,91,124, 151,197,209	**21**,91	37,**58**,84,92,134, 148,157,206	102,197
r=7	-	197	**21**	37,**58**,84,92,134, 148	47
r=8	**2**	**3**,151,167,177	**21**,91	37,92,148,157	**72**
r=9	**2**	**3**	**21**	**58**,92,148,157	37,134,148
r=10	**2**	**3**,113,162,197	**21**,67	37,**58**,92,134,148, 157	-
Replication	EoD_{ss} (99)	EoD_{ss} (121)	EoD_{ss} (188)	EoD_{ss} (190)	EoD_{ss} (213)
r=1	-	**121**	140,176,**188**	**190**	12,47,186,202,**213**
r=2	-	**121**,170,176	**188**	47,84,136,**190**	**213**
r=3	136	**121**,176	**188**	12,47,58,84,136,148, 157,177,184,**190**,197	188,**213**
r=4	213	43,66,79,**121**, 154,157,176	96,140,147,**188**	47,58,84,136,148, 157,**190**,197	**213**
r=5	12,136	66,79,**121**	11,79,108,140,151, 176,**188**,209	84,**190**	12,186,**213**
r=6	136	**121**	151,176,**188**	136,**190**	47,**213**
r=7	136	2,43,58,66,79, **121**,154,157,166, 174,176,206	**188**	12,47,84,136,157, **190**	**213**
r=8	12,136	66,**121**	11,96,121,176,**188**, 209	12,47,84,136,**190**	**213**
r=9	136	66,**121**	-	47,58,84,134,136, 148,**190**,197	16,118,186,**213**
r=10	213	66,79,**121**	72,96,107,118,179, 186,**188**,206	12,37,47,58,84, 134,136,148,157,**190**, 197,206	-

APPENDIX IV

INDIVIDUAL-SPECIFIC MANAGEMENT OF REFERENCE DATA IN ADAPTIVE ENSEMBLES FOR FACE RE-IDENTIFICATION

Miguel De-la-Torre[1,2], Eric Granger[1], Robert Sabourin[1], Dmitry O. Gorodnichy[3]

[1] Laboratoire d'imagerie de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montréal, Canada

[2] Centro Universitario de Los Valles, Universidad de Guadalajara, Ameca, México

[3] Science and Engineering Directorate, Canada Border Services Agency, Ottawa, Canada

Paper submitted to journal "IET - Computer Vision" as an extension of the conference paper (De-la Torre *et al.*, 2013), October 2014

ABSTRACT

In video surveillance, face re-identification allows to recognize target individuals of interest from faces captured across a network of video cameras. In such applications, face recognition is challenging because faces are captured under limited spatial and temporal constraints. Additionally, facial models for recognition are commonly designed using a limited number of representative reference samples from faces captured under specific conditions, regrouped into facial trajectories. Given new reference samples (provided by an operator or through some self-updating process), updating facial models may allow maintaining a high level of performance over time. Although adaptive ensembles have been successfully applied to robust modeling of an individual's facial appearance, reference data samples from a trajectory must be stored for validation. In this paper, a memory management strategy based on Kullback-Leiber (KL) divergence is proposed to rank and select the most relevant validation samples over time in adaptive individual-specific ensembles. When new reference samples become available for an individual, updates to the corresponding ensemble are validated using a mixture of new and previously-stored samples. Only the samples with the highest KL divergence are preserved in memory for future adaptations. This strategy is compared with reference classifiers using videos from the Face in Action data. Simulation results show that the proposed strategy tends

to select discriminative samples from wolf-like individuals for validation. It allows maintaining a high level of performance, while reducing the number of samples per individual by up to 80%.

1. Introduction

In many video surveillance applications, automated face recognition (FR) is increasingly employed to alert a human operator to the presence of individuals of interest appearing in either live (real-time monitoring) or archived (post-event analysis) videos. FR in video surveillance (FRiVS) is employed in a range of applications that involve still-to-video FR (e.g., watchlist screening) and video-to-video FR (e.g., person re-identification). This paper focuses on the problem of re-identifying individuals from faces captured using video surveillance cameras, as found in search and retrieval, face tagging, video summarization and other security-related applications.

Using a decision support system for person re-identification, the operator seeks to captures reference facial trajectories corresponding to a target individual of interest appearing in video feeds, and designs a facial model (*e.g.* templates or statistical representation) to be stored in a gallery. A facial trajectory is defined as a set of facial captures (regions of interest produced by face segmentation) that correspond to the same high quality track of a same individual across consecutive frames. Facial models are typically designed a priori using high quality captures (reference trajectories) obtained under controlled conditions. Then, during operations, facial trajectories captured in live or archived video streams are compared against facial models of individuals enrolled to the system.

Face re-identification in video surveillance is typically performed across a network of surveillance cameras. Accurate and timely responses are required for FR from face trajectories captured in potentially complex semi-constrained (e.g., inspection lane, portal and checkpoint entry) and unconstrained (e.g., cluttered free-flow scene at an airport or casino) environments.

Automated systems require robust operation under a wide variety of conditions, and must be fast and scalable to several enrolments and input videos from several IP cameras.

The unobtrusive capture of video sequences with target individuals provides only a limited amount of high quality reference samples to design facial models. Indeed, faces captured in video surveillance incorporate variations due to pose, illumination, blur, restoration, expression, etc. Updating facial models with new reference target trajectories has been shown to improve or maintain a high level of performance over various capture conditions (Connolly *et al.*, 2012; De-la Torre *et al.*, 2012a, 2014a). Abundant non-target facial trajectories are regrouped in the cohort model (CM, non-target individuals enrolled to the system) and universal model (UM, non-target individuals from operational trajectories). These models provide a source of information for designing discriminant face models, leading the need to select the most relevant samples that avoid biasing matchers towards the negative class (Kubat and Matwin, 1997).

This paper is focused on adaptive video-to-video FR using multi-classifier systems (MCSs). It is assumed that faces captured within trajectories (obtained from post-analysis of video feeds) are used to update facial models. Although adaptive ensembles have previously been applied to face modeling (Connolly *et al.*, 2012; De-la Torre *et al.*, 2012a; Polikar *et al.*, 2001), they require the storage of reference validation samples in a long term memory (LTM) to preserve accuracy. One challenge for practical implementation is bounding the growing number of reference samples collected over several updates. Bounding the size of LTMs raises the issue of selecting the most relevant samples to be preserved in memory to maintain performance (Freni *et al.*, 2008). The selection of the most relevant validation samples, as well as the size of individual-specific LTMs also depends on the specific target individual.

In this paper, a strategy is proposed to select the most representative validation samples for an individual to be stored in a fixed size LTM. It is assumed that an ensemble of 2-class classifiers or detectors per target individual (EoD, target vs. non-target) is used for face matching. When a new reference trajectory becomes available, its target samples extracted from captured regions of interest (ROIs) are combined with non-target samples from the CM and UM selected using

one sided selection (OSS) (Kubat and Matwin, 1997). The corresponding EoD is updated and validated using a mixture of new and pre-stored samples in LTM. Among different relevance measures inspired by techniques in active learning, the Kullback-Leibler (KL) divergence is proposed to accurately rank samples in the overlapping area between target and non-target populations. The least relevant samples are discarded.

The strategy proposed to manage a LTM is evaluated on face trajectories collected in semi-constrained environments from the CMU-FIA database (Goh *et al.*, 2005). Three capture sessions with three months separation are considered for experiments on a scenario with gradual changes, whereas a single capture session with frontal, right and left capture views are considered for a scenario with abrupt changes. For validation, the adaptive MCS is composed of an ensemble of 2-class ARTMAP classifiers for each enrolled individual. Average performance is presented and Doddington zoo (Doddington *et al.*, 1998) analysis is employed to compare individual-specific parameters for LTM management. Using the menagerie terminology introduced in (Li and Wechsler, 2005), this analysis allows to categorize subjects into 4 groups of individuals (sheep, goat, wolf and lamb) according to their performance.

2. Adaptive Face Recognition in Video

Assume that video streams are captured from one or more video cameras. During operations, FRiVS involves several processing steps. First, segmentation isolates the facial regions of interest (ROIs) corresponding to faces appearing in each frame using, e.g., the Viola-Jones algorithm. In order to build face trajectories, a tracker (e.g., CAMSHIFT) simultaneously follows the face of individuals in scene and assigns a same ID to facial ROIs from the same individual. Then, feature extraction extracts and selects discriminant features for classification from the extracted ROIs and arranged into feature vectors. Common feature extraction-selection techniques include the Local Binary Pattern (LBP) algorithm and Principal Component Analysis (PCA). Input feature vectors are compared with facial models, producing matching scores that are compared to individual specific thresholds. In video surveillance applications, the system detects all matching identities where matching scores surpass thresholds. Finally, a decision

fusion allows to combine tracking IDs with the output classifier predictions and accumulate responses over a face trajectory. This process allows for reliable spatio-temporal detection of persons of interest (Matta and Dugelay, 2009).

In literature, matching for FRiVS has been addressed as an open-set problem, where the number of individuals of interest is greatly outnumbered by non-target individuals. Multi-class classifiers have been used in video surveillance with a rejection threshold for unknown individuals. A multi-class classifier designed to address the open set problem in video face recognition is the TCM-kNN (Li and Wechsler, 2005). This matcher takes advantage of transductive inference to generate a class prediction based on randomness deficiency. Modular architectures with a detector (1- or 2-class classifier) per individual have been proposed, allowing to set individual-independent parameters (Jain and Ross, 2002). An individual-specific approach is based on the identification of the decision region(s) in the feature space of individual specific faces, and training a dedicated feed forward neural network for each individual of interest (Kamgar-Parsi *et al.*, 2011). Another example is an SVM-based modular system that was applied to an access control scenario (Ekenel *et al.*, 2009). To improve accuracy and reliability ensembles of 2-class classifiers or detectors (EoD) have been proposed to implement individual-specific detectors. EoDs are co-jointly trained using a dynamic particle swarm optimization (DPSO) based training strategy, generating a diversified pool of ARTMAP neural networks. Trained detectors are selected and combined using boolean combination (BC) (Pagano *et al.*, 2012).

Adaptive systems for FR in video have also been proposed in literature to maintain a high level of performance. These allow to update facial models over time through supervised incremental learning of new data. An incremental learning strategy based on DPSO has been proposed for video-based access control. It allows to evolve an ensemble heterogeneous multi-class classifiers from new data, using a LTM to store validation samples for fitness estimation and to stop training epochs. This approach reduces the effect of knowledge corruption (Connolly *et al.*, 2012). Another adaptive MCS for FRiVS is composed of an ensemble of binary 2-class classifiers per individual, a DPSO module and a LTM. ARTMAP neural networks are used as ensemble members, and the combination function is updated using BC (De-la Torre *et al.*,

2012a). Learn++ is another well-known ensemble-based technique for incremental learning that has been applied to FR. It employs Adaboost to generate a new set of weak classifiers every time new data becomes available, and combines old and new classifiers using weighted majority voting (Polikar *et al.*, 2001).

To assure a high level of accuracy, adaptive MCSs require the storage of reference validation samples in a LTM. However, memory limitations imposed by real-world systems prevent the indefinite growth of the amount of stored validation samples. In literature, editing algorithms like the condensed nearest neighbor have been used to manage a gallery of templates in template matching systems, and bound the amount of reference samples stored in memory (Freni *et al.*, 2008). In this paper adaptive MCSs are considered for FRiVS, where an ensemble of 2-class classifiers is used to estimate the facial model of individuals of interest (De-la Torre *et al.*, 2012a, 2013). An individual-specific strategy is proposed to manage (rank and select) the most informative validation samples over time for each adaptive ensemble.

3. Selection of Representative Samples

Some methods in literature allow to select a subset of representative samples for validation, and the criteria for representativeness is related to the level of information provided for the specific system. Fig. IV-1 presents the levels of selection that are relevant for ensembles of binary 1- or 2-class classifiers.

At the *input data level* **(A)** the dataset itself is used to filter out redundant samples, information about data distributions of samples is not required. At the *classifier level* **(B)** the relevance measure of samples is retrieved from the internal response of the classifiers in the ensemble, to an input sample $\mathbf{a}$. At the *classifier score level* **(C)**, the output scores $S_m^+(\mathbf{a})$ of M classifiers in the ensemble may be combined to produce a measure of relevance. When probabilistic classifiers are used as base classifiers, the computation of relevance measures is based on the combined estimated posterior probability (classification scores S_m^+). At the *classifier decision level* **(D)**, the output predictions $d_m(\mathbf{a})$ of classifiers in the ensemble are combined. Voting strategies can

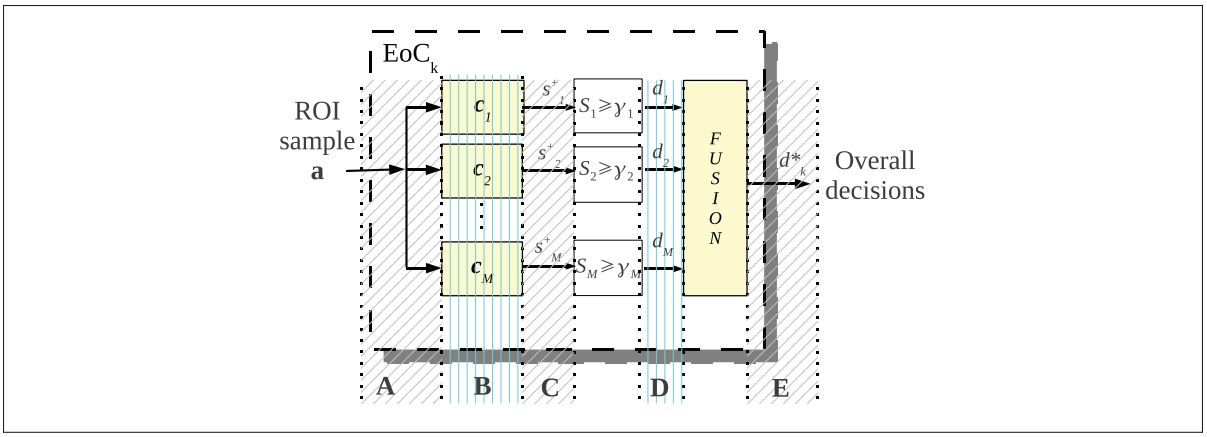


Figure-A IV-1 Levels of ranking that are relevant for an ensemble of detectors (1 or 2-class binary classifiers) for individual k

be used to generate a relevance measure like vote entropy. Finally, at the *ensemble decision level* **(E)**, the global output of the ensemble can be used as a measure of the informativeness of the input sample.

Uninformed Selection. Unlike other levels, methods from level **A** do not require previously trained classifiers to provide information in the selection process. For instance, random under-sampling is the easiest non-heuristic method that randomly eliminates samples from the majority class. Other methods exploit the geometric relationship between samples in feature space, like the condensed nearest neighbor rule (CNN) and one sided selection (OSS) (Guo *et al.*, 2008).

OSS is considered in this paper to select representative samples from the CM and UM. It aims to eliminate the samples from the majority (non-target) that are distant from the decision boundary in the original set D. It starts by building a training set D' with all target samples and one randomly selected non-target sample. Then, 1-NN is trained on D', and used to classify the remaining non-target samples. Misclassified non-target samples are incorporated to D', which at the end will constitute a consistent subset of D.

Informed Selection. Methods at levels **C** and **D** are independent of classification algorithm used in the ensemble as well as combination strategy, and allow to rank and select represen-

tative samples. The only constraint imposed by level **C** lies in the compatibility of scores produced by classifiers, a limitation that can be defeated by using normalization strategies.

A method that operates at level **C** is the *average margin sampling*. It is inspired on the *margin sampling* proposed by Scheffer *et al* in (Scheffer *et al.*, 2001), and is defined as

$$AMS(\mathbf{a}) = \frac{1}{M}\sum_{m}^{M} MS_m(\mathbf{a}) \ , \tag{A IV-1}$$

where M is the number of ensemble members, and $MS_m(\mathbf{a})$ is the margin sampling estimated for each ensemble member c_m given the input sample $\mathbf{a}$. Margin sampling is computed by

$$MS(\mathbf{a}) = S(\omega_{max}, \mathbf{a}) - S(\omega_{2max}, \mathbf{a}) \ , \tag{A IV-2}$$

where $\omega_{max}, \omega_{2max}$ are the first and the second most probable class labels respectively, and $S(\omega)$ is the output score (*e.g.* posterior probability) of a given classifier for class ω. Margin sampling aims to incorporate the posterior probability of the second most likely class label to the relevance measurement.

The disagreement between base classifiers on a test sample $\mathbf{a}$ has also been used as a measure of relevance. For instance, the *Kullback-Leibler* (KL) divergence (or relative entropy), proposed by McCallum and Nigam, operates at level **C** (Kachites McCallum and Nigam, 1998). The KL divergence is defined as

$$KL(\mathbf{a}) = \frac{1}{M}\sum_{m=1}^{M}\left(\sum_{i\in\Omega} S_m^i(\mathbf{a}) \log \frac{S_m^i(\mathbf{a})}{\hat{P}^i_{EoD_k}(\mathbf{a})}\right) \ , \tag{A IV-3}$$

where M is the number of classifiers in the ensemble, and $\hat{P}^i_{EoD_k}(\mathbf{a})$ given by Eqn. A IV-4 is the consensus probability that the class $i \in \Omega$ is the correct label for sample $\mathbf{a}$, given the scores $S_n^i(\mathbf{a})$ produced by the base classifiers.

$$\hat{P}^i_{EoD_k}(\mathbf{a}) = \frac{1}{M}\sum_{n=1}^{M} S_n^i(\mathbf{a}) \ . \tag{A IV-4}$$

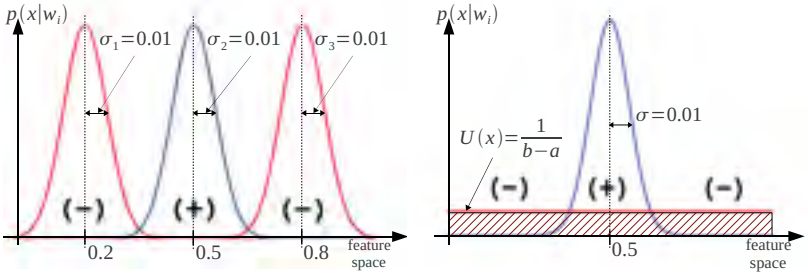
For KL divergence, the most informative samples are those with the largest average difference between the class distributions of any one of the committee members and the consensus.

An example of level **D** relevance measure is the *vote entropy* (Dagan and Engelson, 1995), defined as

$$VE(\mathbf{a}) = -\sum_{i \in \Omega} \frac{V(\omega_i, \mathbf{a})}{M} \log \frac{V(\omega_i, \mathbf{a})}{M} \quad , \tag{A IV-5}$$

where $V(\omega_i, \mathbf{a})$ is the number of votes for the class $\omega_i \in \Omega$ provided by the ensemble. Similarly to KL divergence, VE increases with the disagreement in the ensemble members, but its resolution (*e.g.*, ranking levels) is bounded by the number of base classifiers in the ensemble.

Synthetic Analysis. For more insight on the selective capacity of the relevance measures, two synthetic 2-class problems were designed in the 1D space. Fig. IV-2 shows the original probability distributions of data. Central Gaussian distribution in Fig. IV-2a and IV-2b have a center of mass $\mu_2 = 0.5$. Centers of mass of the non-target distributions in Fig. IV-2a are $\mu_1 = 0.2$ and $\mu_3 = 0.8$, and in Fig. IV-2b the non-target samples are randomly drawn according to a uniform distribution. All Gaussians have a variance of $\sigma = 0.01$.



(a) Problem 1 **(b)** Problem 2

Figure-A IV-2 Data distributions used to generate the training data for both problems. Central Gaussian distributions in both figures generate the positive (+) samples, and left and right distributions generate negative (-) samples

A pool of 7 probabilistic Fuzzy ARTMAP (PFAM) classifiers was trained for each problem using balanced data. The PFAM classifier combines the Fuzzy ARTMAP learning to encode category prototypes and update centers of mass of estimated class distributions (Lim and Harrison, 1997). A DPSO learning strategy was used for base classifiers generation and hyperparameter optimization (Connolly *et al.*, 2012).

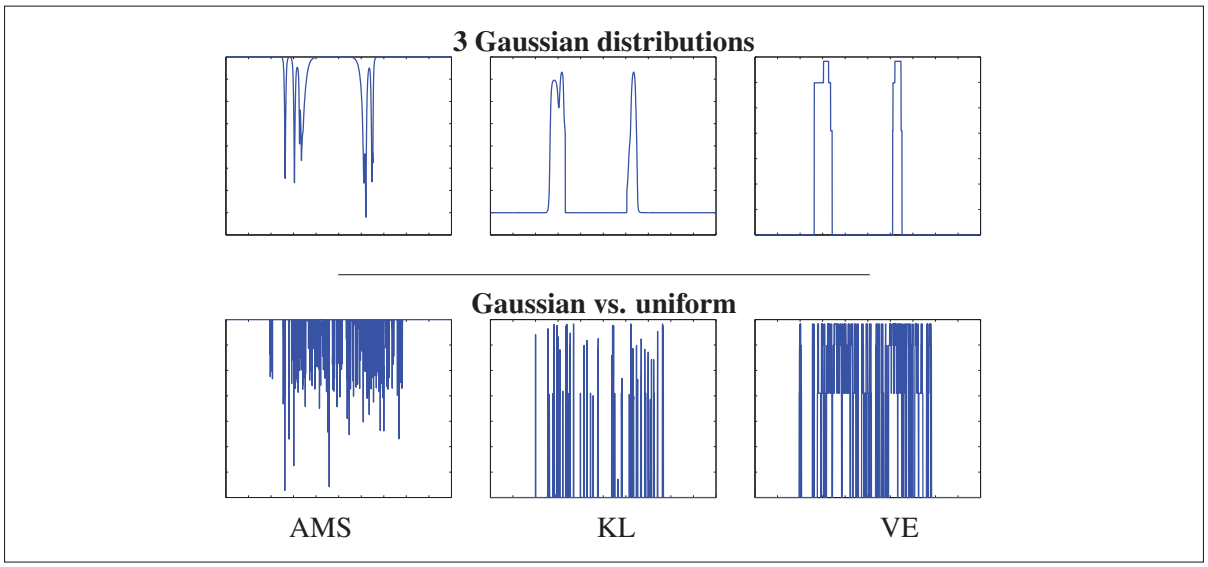


Figure-A IV-3 Value of relevance measures obtained over the feature space with an EoD (PFAM) for the 3 Gaussians (top) and Gaussian vs. uniform (bottom) problems

The value of relevance measures produced by the ensembles are presented on Fig. IV-3. The three measures show a good characterization of the overlapping region between target and non-target populations, specially on the problem with three Gaussians. Vote Entropy shows a lower resolution than KL divergence and AMS, and the smoothness of the KL divergence curve shows a better representation of the overlapping area. In this paper, the KL divergence is employed to implement a strategy to assess the relevance of reference samples to manage a fixed size memory.

4. Individual-Specific Management of LTM

Fig. IV-4 presents the modular architecture for FRiVS that allows for supervised adaptation of facial models from new trajectories. During operations, the system will process the ROI patterns extracted from each frame, and along input trajectories. ROI feature vectors are extracted and presented to each EoD_k. Using a face tracking algorithm, different faces in a video sequence are followed frame to frame and regrouped, and the successive predictions p_k from EoD_k for each trajectory are accumulated over time for spatio-temporal recognition, in order to provide an overall prediction for each track ID. Finally, an individual specific threshold is applied to the accumulation curves of each EoD_k in order to generate an overall decision d_k for each EoD_k. Note that there are several accumulation modules per track ID, to simultaneously recognize several people at a time in the scene.

During design/update, each EoD_k performs independent supervised incremental learning. When a new trajectory T_k becomes available for a person k, OSS is used to form a consistent individual-specific training set D_k with all target samples and non-target samples selected from CM and UM. Then, a DPSO-based strategy is employed to generate a new pool of diversified binary classifiers that are combined with previously trained detectors corresponding to person k (De-la Torre *et al.*, 2012a). A fixed size LTM is maintained with validation samples that are representative of the overlapping zone between target and non-target distributions. The KL divergence measure (Eq. A IV-3) is employed to rank reference samples and store the λ_k most representative in the LTM, where λ_k is the size of the LTM for person k enrolled to the system. At each adaptation step, new validation samples are combined with those stored in the LTM to accurately estimate a new fusion function and select an operations point.

Algorithm 4.1 shows the procedure followed by the management strategy to rank and select representative validation samples to be stored in the LTM_k. When a new validation set D with target and non-target samples becomes available for individual k, all samples are ranked according to the KL divergence. Then, the $\lambda_k/2$ highest ranked target samples, as well as the $\lambda_k/2$ highest ranked non-target samples are preserved, whereas the rest are discarded.

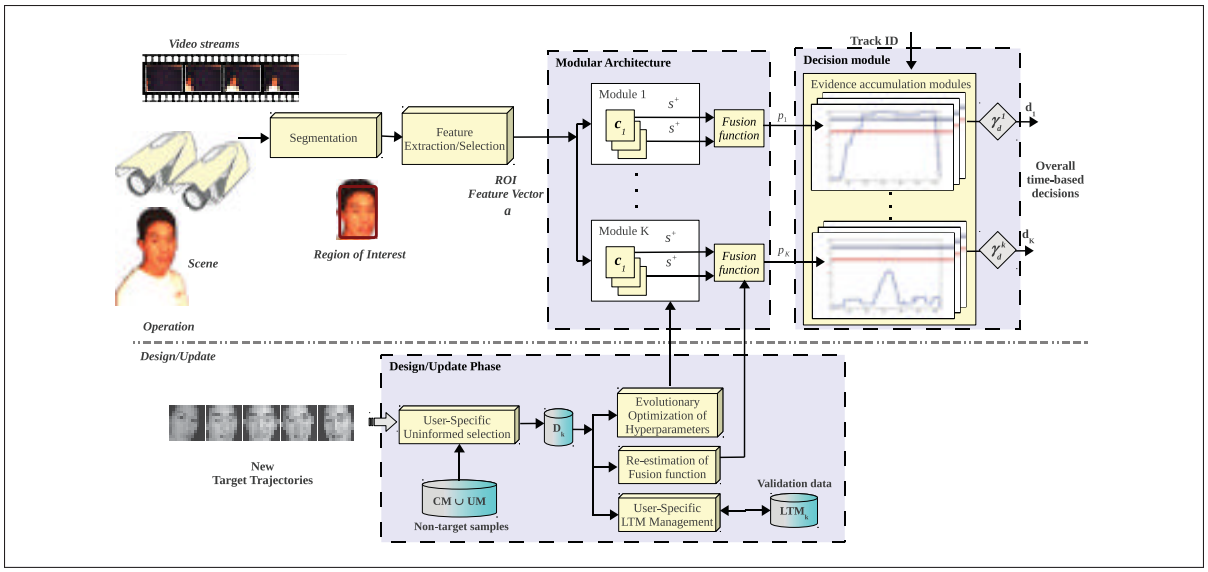


Figure-A IV-4 Adaptive MCS for FRiVS. In the design/update phase, when a new face trajectory T_k becomes available for a person k, a training set D_k is formed with all its target samples, and non-target samples selected from CM and UM using OSS. Then, an evolutionary optimization strategy is employed to generate a new pool of diversified classifiers with optimized hyper parameters, and the decision-level fusion function is updated based on new data and pre-stored reference samples (from the LTM). Finally the λ_k most relevant samples from previous and newly-learned trajectories are stored in LTM according to the KL divergence

Algorithm 4.1: KL relevance subsampling for the EoD$_k$

Input : D, $S_k(a_i)$, λ_k // Validation data, scores and size of LTM_k

Output : Dr // Representative samples

for $a_i \in D$ **do**

  $r_i = KL(S_k(a_i))$ // Rank with Eq. A IV-3

$D \Leftarrow sort(D, r, d)$ // Sort D according to r_i

$Dr^+ \Leftarrow first\_pos(D, \lceil \frac{\lambda_k}{2} \rceil)$

$Dr^- \Leftarrow first\_neg(D, \lceil \frac{\lambda_k}{2} \rceil)$

$Dr \Leftarrow Dr^+ \cup Dr^-$

The new set Dr is formed from old and new validation samples that are difficult to classify by old and new classifiers. Then, the selection is based on past and present information retrieved from the classifiers by choosing the samples in the overlapping area of the target and non-target

distributions. Thus, the proposed selection strategy allows to store the samples that contain the most relevant information to define the decision frontier.

5. Experimental Methodology

The CMU Face in Action (FIA) database (Goh *et al.*, 2005) is employed to characterize the proposed strategy in a person re-identification scenario that presents gradual and abrupt changes. The FIA database consists of 20 second videos of face data from 180 participants mimicking a passport checking scenario. An array of 6 cameras horizontally positioned at the face level capture the scene at 30 fps. Pairs of cameras were positioned at 0^o (frontal) and $\pm 72^o$ (left and right) angle with respect to the individual. Three cameras were set to an 8-mm focal-length (zoomed), resulting in face areas around 300×300 pixels, and the other three to a 4-mm focal-length (unzoomed) resulting in face areas around 100×100 pixels. The cameras utilize the Sony ICX424 sensor, with a maximum resolution of 640x480 pixels and a 6mm diagonal image size. Data has been captured on three sessions separated by a three months interval for each individual.

Facial trajectories were formed with facial regions segmented using the Viola-Jones algorithm (Viola and Jones, 2004) (see Fig. IV-5). An ideal face tracker is assumed, and all images were scaled to the resolution of the smallest face obtained after face detection (70x70 pixels). The Multi Scale LBP (Ojala *et al.*, 2002) feature extractor has been used with three different block sizes (3×3, 5×5 and 9×9), along with pixel intensities features. Resulting features were combined into feature vectors, and PCA was applied to select the 32 most discriminant projected features.

Ten individuals were randomly selected for re-identification, and one EoD_k was designed for each. 88 of the remaining individuals are selected as part of the universal model (UM), and the rest are considered as never seen test individuals. The cohort model (CM) comprises trajectories from non-target individuals enrolled to the system. It is important to highlight that individuals from the UM never appear in test. Face trajectories from individuals of interest

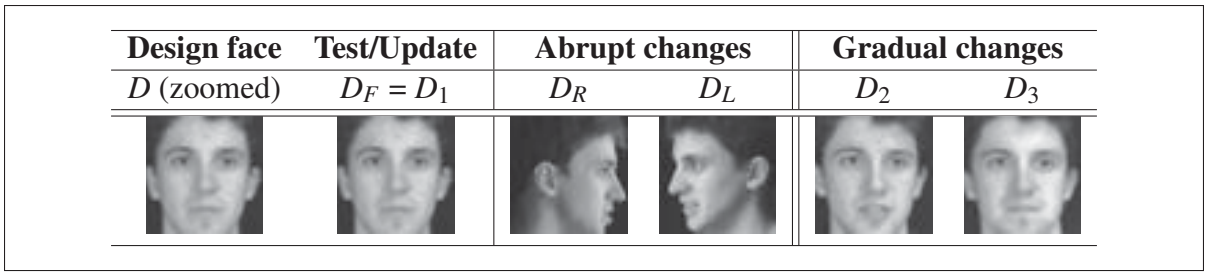


Figure-A IV-5 Samples of design/update facial regions from one of the individuals enrolled to the system (ID 188). Faces were detected in video sequences from the FIA database using the Viola-Jones face detector trained with frontal faces for gradual changes, and frontal, right and left poses for abrupt changes

contain between 80 and 239 facial regions, and non-target training and test samples differ in each dataset.

Prior to computer simulations, five data subsets have been prepared. Trajectories in the design dataset D are comprised of target ROI patterns from the the zoomed view of capture session 1. In order to build a scenario with gradual changes (age), the test/adaptation datasets D_1 to D_3 have been constructed with ROI patterns from the unzoomed view of capture sessions 1 to 3 respectively. On the other hand, for the scenario with abrupt changes (pose), the test/adaptation datasets D_F, D_R and D_L have been constructed with ROI patterns from the unzoomed view of capture session 1, with the frontal, right and left cameras respectively. Non-target samples are independently selected for each of the training/validation sets picked from the CM and UM, using OSS (Kubat and Matwin, 1997).

The classifiers were initially trained using trajectories in the design set D, and tested on trajectories in D_1 (or equivalently D_F for the scenario with abrupt changes), obtaining the performance for the first evaluation. After performance evaluation on D_1 (D_F) the classifiers were updated with trajectories in D_1 (D_F) and tested on D_2 (D_R). The same process was repeated for update/test on D_2 (D_R) and D_3 (D_L) respectively in both scenarios with gradual and abrupt changes.

The approaches capable of incremental learning (PFAM, Learn++ (PFAM) and EoD$_k$ (PFAM)) were updated with only the new labeled dataset. In contrast, TCM-kNN was trained on batch mode, learning from scratch the previous and new samples. The MCS used for LTM analysis was composed of an ensemble of 2-class Probabilistic Fuzzy ARTMAP (PFAM) classifiers per individual, EoD$_k$ (PFAM). The DPSO learning strategy was used for classifiers generation and hyperparameters optimization, and BC was applied for decision level fusion of classifiers on the ROC space (De-la Torre *et al.*, 2012a). The LTM was managed according to the KL divergence with six individual-specific values of λ_k were explored: 0, 25, 50, 75, 100 and ∞.

Evaluation was performed following 2×5-fold cross-validation for 10 independent trials. Target samples from the learning set were randomly split according to a uniform distribution, in 5 folds of the same size. The folds were first distributed in three different design sets, including two folds for training (D_t^t), $1\frac{1}{2}$ folds to stop training epochs (D_t^e), and $1\frac{1}{2}$ folds for fitness evaluation (D_t^f). Once the classifiers were trained, D_t^e and D_t^f are combined, randomized and divided in two equally distributed subsets to produce a validation data for threshold/fusion function estimation (D_t^c), and to select the operations point (D_t^s). Each fold was assigned to a different training/validation set at each replica of the experiment. At replication 5, the five folds were regenerated after a randomization of the sample order for each class, and the process was repeated to generate a standard error on ten different assignments.

Reference approaches in comparison include TCM-kNN, single PFAM in incremental learning mode and Learn++ with 7 PFAM base classifiers. TCM-kNN was trained with a fixed $k = 1$ on a batch learning scheme. PFAM classifiers used in all other approaches, were trained using DPSO based learning strategy to optimize hyperparameters. Validating the number of training epochs for classifier convergence was performed on D_t^e, whereas particle fitness was evaluated on D_t^f. The DPSO algorithm was initialized with a swarm of 60 particles, and a maximum of 5 particles within each of the 6 subswarms. The algorithm was set to run a maximum of 30 iterations, allowing 5 extra iterations to ensure convergence. Once the global best particle is found, its classifier and the 6 local bests from each subswarm were added to the EoD.

6. Simulation Results

Figure IV-6 presents the average performance of the system for the 10 individuals of interest, after incremental learning. The ROC and PROC performance spaces are used for comparison, with partial area under the ROC curve for $0 \leq fpr \leq 0.05$ (pAUC (5%)), and empiric estimation of tpr, fpr and F_1 measure.

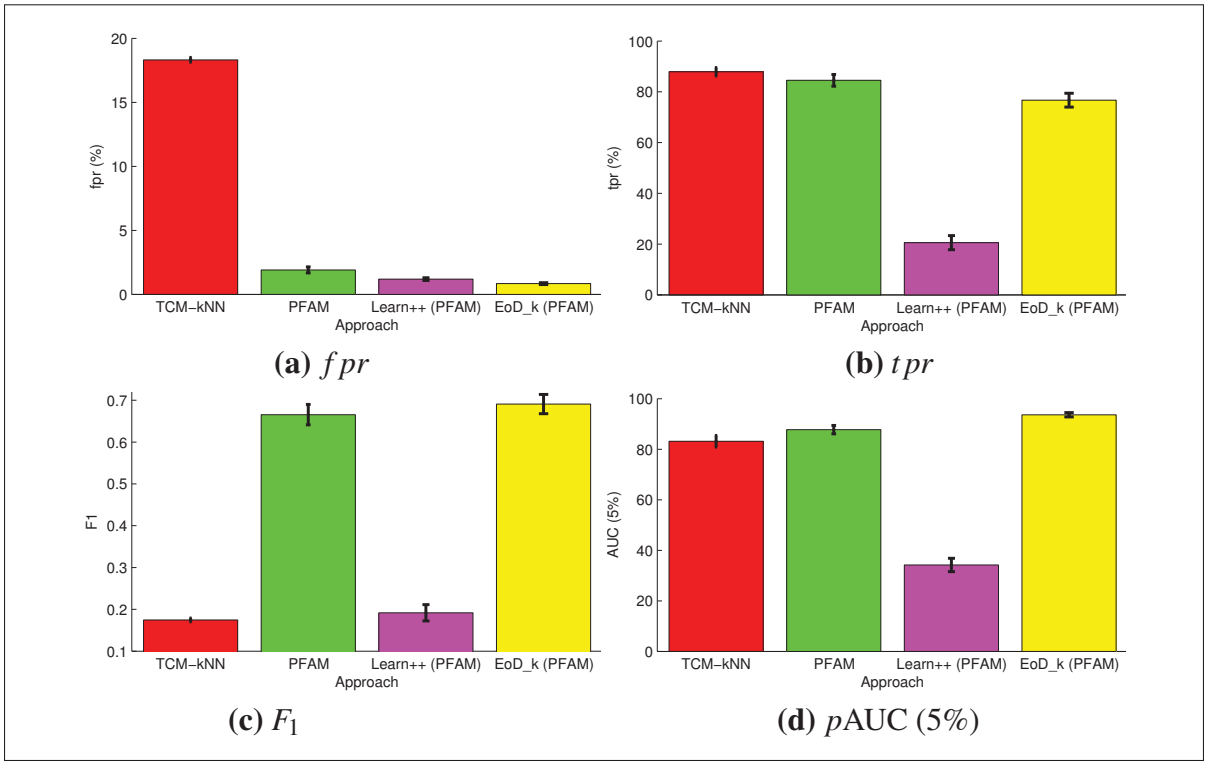


Figure-A IV-6 Average transaction-based performance of the different classifiers after two updates ($D_1 \rightarrow D_2 \rightarrow D_3$). More details on this comparison can be found in (De-la Torre *et al.*, 2013). The tpr, fpr and F_1 measure are estimated at the operations point selected for a fixed $fpr = 1\%$

Figure IV-6 **(a)** shows that TCM-kNN yields the highest fpr, that is related to the difficulty faced by multi-class classifiers in finding multiple boundaries during the same optimization process. In contrast, Learn++ (PFAM) and EoD$_k$ (PFAM) LTM$_{kl,\lambda_k}$ present the lowest fpr, proving the enhanced capacity of ensemble-based classifiers to discard non-target samples.

Besides, Fig. IV-6 **(b)** shows that TCM-kNN presents the highest tpr, followed by the PFAM and EoD$_k$ (PFAM) in the third place. In general, Figures IV-6 **(c)** and **(d)** show that the EoD$_k$ (PFAM) with a LTM managed with KL divergence presents the highest overall performance with a lower standard error.

Table IV-1 presents the average performance obtained after incremental learning in the scenarios with gradual and abrupt changes. Regarding the pAUC (5%), the tendency shown by the system in a scenario with gradual changes is characterized by an increase in the performance after two adaptations. An opposite tendency is shown on a scenario with abrupt changes, where the performance is constantly decreasing. This tendency is natural since facial models are designed with frontal faces, and it is required to recognize the individuals on right or left poses (see Fig. IV-5). However, the system behaves differently for each individual in each scenario, and the impact of using a LTM is also different in each case.

Table-A IV-1 Average performance of the system on 10 individuals and 10 trials, for the scenarios with gradual (top) and abrupt (bottom) changes. The operations point was selected at $fpr = 1\%$

fpr (%) ↓	tpr (%) ↑	F_1 ↑	pAUC (5%) ↑
Gradual changes ($D_1 \rightarrow D_2 \rightarrow D_3$)			
EoD$_k$ (PFAM) $LTM_{KL,\lambda_k=\infty}$			
0.62 ±0.09 → 0.67 ±0.05 → 0.84 ±0.07	77.02 ±2.10 → 45.51 ±3.63 → 76.70 ±2.71	0.6789 ±0.0177 → 0.4041 ±0.0308 → 0.6909 ±0.0231	92.88 ±0.81 → 72.03 ±2.76 → 93.64 ±0.84
Abrupt changes ($D_F \rightarrow D_R \rightarrow D_L$)			
EoD$_k$ (PFAM) $LTM_{KL,\lambda_k=\infty}$			
0.62 ±0.09 → 5.38 ±1.13 → 2.73 ±0.34	77.02 ±2.10 → 13.48 ±2.444 → 11.68 ±2.42	0.6789 ±0.0177 → 0.0571 ±0.0121 → 0.0605 ±0.0147	92.88 ±0.81 → 22.0747 ±2.598 → 19.68 ±2.5450

Table IV-2 presents the performance of the ensemble during incremental learning for two individuals, using $\lambda_k = 25, 75$ and 100. EoD$_{58}$ was selected because of its good initial performance (pAUC (5%) $\geq 95\%$). This individual is easy to detect by the system ($tpr > 80\%$), and easy to differentiate against non-target individuals ($fpr < 1\%$) – it is a *sheep*-like subject in the Doddington zoo taxonomy (Li and Wechsler, 2005). Conversely, EoD$_{188}$ was selected because of its low initial performance (pAUC (5%) $< 95\%$). It corresponds to a *lamb*-like individual that even though is easy to detect by the system ($tpr > 80\%$), it is also easy to imperson-

ate ($fpr > 1\%$). For individual 188, the test on D_1 throws 32 non-target individuals that are wrongly detected more than 1% of the time (*wolves*).

Table-A IV-2 Average performance of the EoD_{58} and EoD_{188} after tests on scenarios of gradual ($D_1 \rightarrow D_2 \rightarrow D_3$) and abrupt ($D_F \rightarrow D_R \rightarrow D_L$) changes

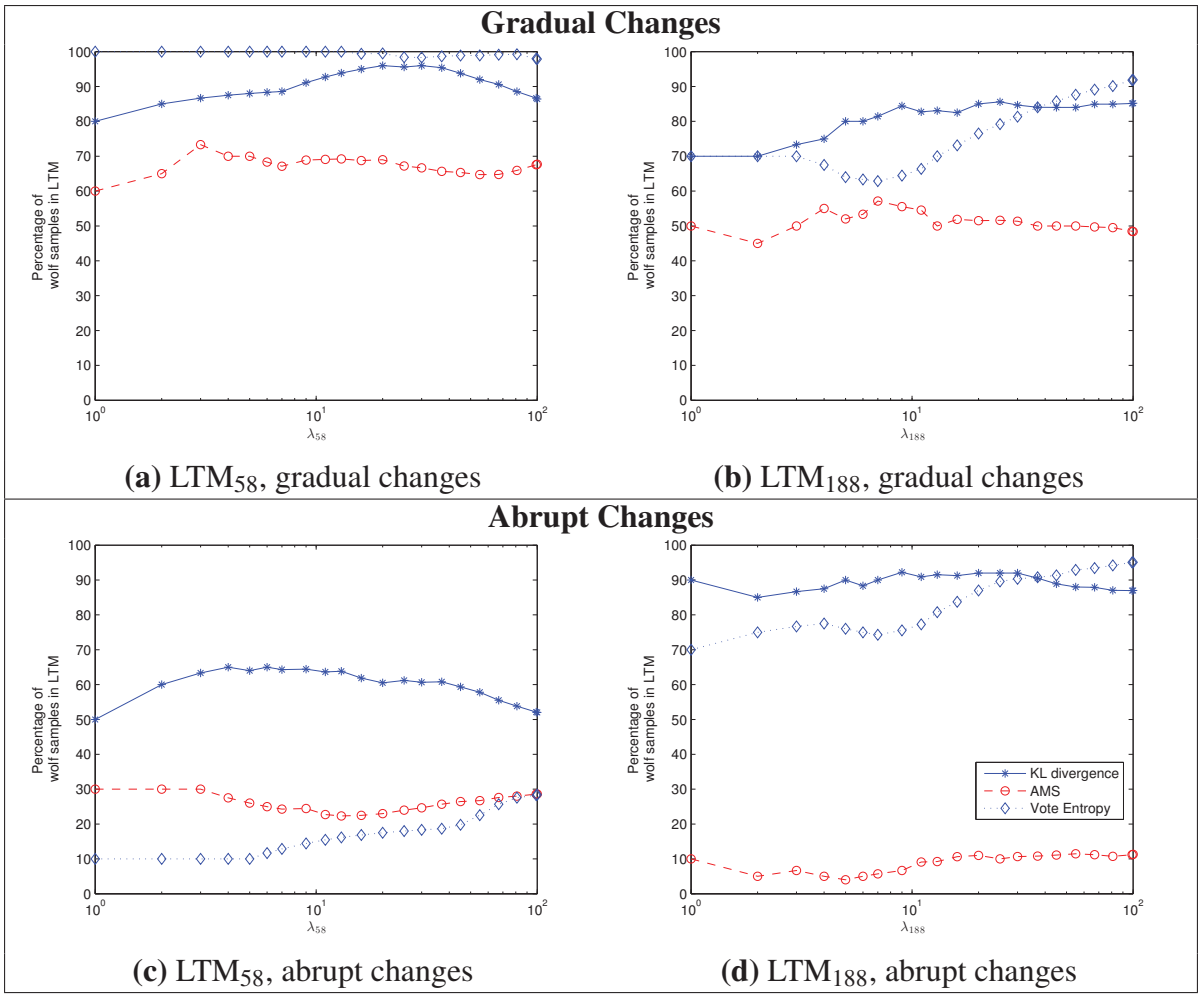
	Gradual changes		**Abrupt changes**	
	EoD_{58}	**EoD_{188}**	**EoD_{58}**	**EoD_{188}**
$\text{LTM}_{KL,\lambda_k=25}$				
$fpr \downarrow$	0.23 ±0.09 → 0.87 ±0.07 → 3.92 ±0.71	2.54 ±0.57 → 1.01 ±0.10 → **0.84 ±0.24**	0.23 ±0.09 → 29.51 ±1.83 → 3.71 ±0.407	2.54 ±0.57 → 1.952 ±0.17 → 3.17 ±0.64
$tpr \uparrow$	84.43 ±3.33 → 39.49 ±7.01 → 90.93 ±3.02	89.58 ±4.26 → 84.88 ±5.36 → 97.29 ±0.82	84.43 ±3.33 → 43.33 ±3.35 → 0.62 ±0.15	89.58 ±4.26 → 28.33 ±2.05 → **6.15 ±0.87**
$F_1 \uparrow$	0.8492 ±0.023 → 0.4029 ±0.061 → 0.5710 ±0.043	0.4720 ±0.054 → 0.6594 ±0.038 → **0.8730 ±0.027**	0.8492 ±0.023 → 0.0134 ±0.001 → 0.0016 ±0.001	0.4720 ±0.054 → 0.3119 ±0.021 → 0.0370 ±0.005
pAUC (5%) $\uparrow$	98.45 ±0.23 → 72.46 ±3.74 → 97.18 ±1.09	91.12 ±2.41 → 96.43 ±0.80 → 99.64 ±0.07	98.45 ±0.23 → 8.15 ±0.57 → 8.93 ±0.4281	91.12 ±2.41 → 38.71 ±1.73 → 14.51 ±0.97
$\text{LTM}_{KL,\lambda_k=75}$				
$fpr \downarrow$	0.23 ±0.09 → 0.84 ±0.10 → 4.29 ±0.62	2.54 ±0.57 → 1.02 ±0.10 → 1.07 ±0.31	0.23 ±0.09 → 33.23 ±1.71 → **2.98 ±0.13**	2.54 ±0.57 → 2.62 ±0.16 → 1.83 ±0.29
$tpr \uparrow$	84.43 ±3.33 → 41.49 ±7.76 → 94.65 ±3.25	89.58 ±4.26 → 89.53 ±3.21 → **97.60 ±0.64**	84.43 ±3.33 → 48.33 ±3.96 → 0.16 ±0.049	89.58 ±4.26 → 26.51 ±1.86 → 5.38 ±1.11
$F_1 \uparrow$	0.8492 ±0.023 → 0.4171 ±0.064 → 0.5619 ±0.053	0.4720 ±0.054 → 0.6838 ±0.026 → 0.8511 ±0.033	0.8492 ±0.023 → 0.0122 ±0.001 → 0.0007 ±0.001	0.4720 ±0.054 → 0.2743 ±0.017 → **0.0385 ±0.007**
pAUC (5%) $\uparrow$	98.45 ±0.23 → 71.92 ±3.50 → **98.60 ±0.77**	91.12 ±2.41 → 96.21 ±0.67 → 99.63 ±0.09	98.45 ±0.23 → 8.44 ±0.60 → **9.78 ±0.45**	91.12 ±2.41 → 38.19 ±1.22 → 17.94 ±1.16
$\text{LTM}_{KL,\lambda_k=100}$				
$fpr \downarrow$	0.23 ±0.09 → 0.84 ±0.08 → **3.64 ±0.73**	2.54 ±0.57 → 1.09 ±0.14 → **0.84 ±0.19**	0.23 ±0.09 → 30.42 ±1.56 → 4.14 ±0.23	2.54 ±0.57 → 2.59 ±0.22 → **1.74 ±0.26**
$tpr \uparrow$	84.43 ±3.33 → 38.28 ±8.46 → **95.81 ±1.63**	89.58 ±4.26 → 88.08 ±3.06 → **97.60 ±0.52**	84.43 ±3.33 → 45.00 ±3.52 → **4.06 ±0.88**	89.58 ±4.26 → 31.52 ±2.08 → 5.38 ±1.01
$F_1 \uparrow$	0.8492 ±0.023 → 0.3808 ±0.071 → **0.6168 ±0.053**	0.4720 ±0.054 → 0.6669 ±0.032 → 0.8720 ±0.022	0.8492 ±0.023 → 0.0125 ±0.001 → **0.0151 ±0.004**	0.4720 ±0.054 → 0.3231 ±0.020 → 0.0379 ±0.007
pAUC (5%) $\uparrow$	98.45 ±0.23 → 71.91 ±3.56 → 98.36 ±0.79	91.12 ±2.41 → 96.25 ±0.55 → **99.67 ±0.09**	98.45 ±0.23 → 8.44 ±0.61 → 9.33 ±0.47	91.12 ±2.41 → 41.25 ±1.20 → **19.42 ±1.13**

Regarding the scenario with gradual changes, the F_1 measure for EoD_{58} after test on D_2, results show a performance that declines more importantly for EoD_{58} with $\lambda_{58} = 100$, and using a $\lambda_{58} = 75$ shows the best performance. However, after test on D_3, the appearance of new representative samples in the LTM leads to a recovery in the performance. A similar but smaller recovery is presented by EoD_{58} in the scenario with abrupt changes, suggesting that sheep-like individuals benefit from high λ_k values either in scenarios with gradual or abrupt changes.

A different trend is shown by EoD_{188} in the scenario with gradual changes, which in general presents a performance increase every time it is updated, regardless the value of λ_{188}. A comparison between λ_{188} values shows that there is no significant difference between using a large or small LTM, indicating that the performance of the EoD_{188} for this lamb-like individual is maintained using this KL-based selection, even with small λ_{188} values (e.g. $\lambda_{188} = 25$). Note

that the average number of samples selected by OSS for validation in experiments is 139.1 $\pm$ 5.07 (global average for the 10 individuals over the 10 trials), and $\lambda_{188} = 25$ samples constitutes the 17.97% of the data.

Regarding the scenario with abrupt changes, the EoD_{188} shows a performance decrease as expected by pose changes. But regarding its final performance, the use of large λ_{188} values significantly benefits its final performance. This suggests that lamb-like individuals are benefited by large λ values in scenarios with abrupt changes.



(a) LTM_{58}, gradual changes

(b) LTM_{188}, gradual changes

(c) LTM_{58}, abrupt changes

(d) LTM_{188}, abrupt changes

Figure-A IV-7 Average percentage of samples from wolf-like individuals for the EoD_{58} (a and c) and EoD_{188} (b and d), in the scenarios with gradual (upper graphs) and abrupt (lower graphs) changes

Samples from wolf-like individuals degrade the fpr of EoDs for lamb-like individuals, and are useful for system's validation, allowing for better discrimination. Fig. IV-7 shows the percentage of samples from wolf-like individuals selected by the KL algorithm for the EoD_{58} and EoD_{188}, using a λ_k that grows up to 100 samples characterizing the scenarios with gradual and abrupt changes. The three selection strategies presented in section 3 are compared. Regarding the scenario with gradual changes, it can be seen that LTM management strategies based on KL divergence and VE are successful in storing samples from wolf-like individuals, and the KL divergence retrieves the highest percentage for the lamb-like individual 188 (Fig. IV-7b). Results for the scenario with abrupt changes reveal that the KL divergence overcomes the other strategies at retrieving a greater proportion of samples from wolf-like individuals, either for lamb- or sheep-like target individuals . This becomes more evident for small values of λ.

Finally, when a new trajectory for an individual of interest becomes available, it takes around 150 min. to update its facial model, and the modular architecture allows for parallel update of multiple facial models. The algorithm was implemented in Matlab® R2010B, running on Linux Gentoo, on a 2.53GHz Intel® Xeon® processor. This makes the system appropriate for off-line update from, e.g., daily police reports.

7. Conclusion

In this paper, an individual-specific strategy was proposed for the management of reference samples used for validation of adaptive ensembles applied to face re-identification. When new reference samples become available for an individual enrolled to the system, its facial regions are combined with non-target samples from the universal and cohort models selected with OSS. Old and new validation samples are combined and ranked using Kullback-Leibler divergence, and the highest ranked are stored in a LTM for future validations. The theoretical foundation of this relevance measure lies on the relative entropy, where the disagreement between ensemble members is an indicator of the informativeness of reference samples.

This strategy was tested on real-world CMU-FIA video data emulating scenarios with gradual (aging) and abrupt (pose) changes in the classification environment. Simulation results indicate that using the proposed strategy allows individual-specific ensembles to maintain a level of performance comparable to that achieved by an ensemble where all validation samples are preserved, yet storing less than 20% of samples. Comparing different LTM sizes (λ_k) for individual-specific ensembles suggests that sheep-like individuals benefit from high λ_k values, whereas low λ_k values may be selected for lamb-like individuals. This is related to the capacity of the KL divergence to rank and select samples from wolf-like individuals, compared to vote entropy and average margin sampling. Future research includes investigating strategies to find the optimal amount of samples required for each EoD, affecting a trade-off between performance and resources.

BIBLIOGRAPHY

Ahmad, I., Z. He, M. Liao, F. Pereira, and M. T. Sun. 2008. "Special issue on Video Surveillance". *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, n° 8, p. 1001-1005.

Amis, Gregory P. and Gail A. Carpenter. 2007. "Default ARTMAP 2". In *IEEE International Conference on Neural Networks - Conference Proceedings*. (Orlando, FL, United states 2007), p. 777 - 782.

Anagnostopoulos, G.C. and M. Georgiopoulos. 2001. "Ellipsoid ART and ARTMAP for incremental clustering and classification". In *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*. p. 1221 -1226.

Anagnostopoulos, G.C. and M. Georgiopulos. 2000. "Hypersphere ART and ARTMAP for unsupervised and supervised, incremental learning". In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*. p. 59 -64 vol.6.

Arandjelovic, Ognjen and Roberto Cipolla. 2006. "Incremental learning of temporally-coherent Gaussian mixture models". In *Technical Paper - Society of Manufacturing Engineers*.

Barandela, R., R.M. Valdovinos, and J.S. Sánchez. 2003. "New Applications of Ensembles of Classifiers". *Pattern Analysis & Applications*, vol. 6, n° 3, p. 245-256.

Barreno, Marco, Alvaro A. Cardenas, and J. D. Tygar. 2008. "Optimal ROC curve for a combination of classifiers". In *In Advances in Neural Information Processing Systems (NIPS)*. p. 1-8.

Barry, M. and E. Granger. 2007. "Face recognition in video using a What-and-Where fusion neural network". In *IEEE International Conference on Neural Networks - Conference Proceedings*. (Orlando, FL, United states 2007), p. 2256 - 2261.

Bartlett, Marian Stewart, Javier R. Movellan, and Terrence J. Sejnowski. 2002. "Face recognition by independent component analysis". *IEEE Transactions on Neural Networks*, vol. 13, n° 6, p. 1450-1464.

Belhumeur, P.N., J.P. Hespanha, and D.J. Kriegman. 1997. "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, n° 7, p. 711 - 20.

Bella, Antonio, Cesar Ferri, Jose Hernandez-Orallo, and Maria Jose Ramirez-Quintana. 2010. "Quantification via probability estimators". (Sydney, NSW, Australia 2010), p. 737 - 742.

Bengio, S. and J. Mariéthoz. 2007. "Biometric Person Authentication IS A Multiple Classifier Problem". In *7th International Workshop on Multiple Classifier Systems*. p. 513-522.

Best-Rowden, L., B. Klare, J. Klontz, and A.K. Jain. Sept 2013. "Video-to-video face matching: Establishing a baseline for unconstrained face recognition". In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*. p. 1-8.

Bradski, Gary R. 1998. "Computer Vision Face Tracking For Use in a Perceptual User Interface". *Intel Technology Journal*, vol. Q2, p. 1-15.

Brew, A. and P. Cunningham. 2009. "Combining cohort and UBM models in open set speaker identification". In *2009 Seventh International Workshop on Content-Based Multimedia Indexing (CBMI)*. (Piscataway, NJ, USA 2009), p. 62 - 7.

Brew, Anthony and Padraig Cunningham. 2010. "Combining cohort and UBM models in open set speaker detection". In *Proceedings on Multimedia Tools and Applications*. (Van Godewijckstraat 30, Dordrecht, 3311 GZ, Netherlands 2010), p. 141 - 159.

Britto, Alceu S., Robert Sabourin, and Luiz E.S. Oliveira. 2014. "Dynamic selection of classifiers – A comprehensive review". *Pattern Recognition*, vol. 47, p. 3665-3680.

Burghouts, G.J., K. Schutte, H. Bouma, and R.J.M. Hollander. 2014. "Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos". *Machine Vision and Applications*, vol. 25, n° 1, p. 85-98.

Carpenter, Gail A. and Natalya Markuzon. 1998. "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases". *Neural Networks*, vol. 11, n° 2, p. 323 - 336.

Carpenter, Gail A. and Boriana L. Milenova. 1999. "Distributed ARTMAP". In *Proceedings of the International Joint Conference on Neural Networks*. (Washington, DC, USA 1999), p. 1983 - 1987.

Carpenter, Gail A. and William D. Ross. 1995. "ART-EMAP: a neural network architecture for object recognition by evidence accumulation". *IEEE Transactions on Neural Networks*, vol. 6, n° 4, p. 805 - 818.

Carpenter, Gail A., Stephen Grossberg, and John H. Reynolds. 1991. "ARTMAP. Supervised real-time learning and classification of nonstationary data by a self-organizing neural network". *Neural Networks*, vol. 4, n° 5, p. 565-588.

Carpenter, Gail A., Stephen Grossberg, Natalya Markuzon, John H. Reynolds, and David B. Rosen. 1992. "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps". *IEEE Transactions on Neural Networks*, vol. 3, p. 698-713.

Chan, Yee Seng and Hwee Tou Ng. 2006. "Estimating Class Priors in Domain Adaptation for Word Sense Disambiguation". In *Proceedings of the 21st International Conference on*

Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. (Stroudsburg, PA, USA 2006), p. 89–96.

Chen, Li-Fen, H.-Y.M. Liao, and Ja-Chen Lin. 2001. "Person identification using facial motion". In *Image Processing, 2001. Proceedings. 2001 International Conference on*. p. 677-680 vol.2.

Chen, Y-C., V. M. Patel, P. J. Phillips, and R. Chellappa. (submit Oct. 2013) 2014. "Dictionary-based face and person recognition from unconstrained video". *IEEE Transactions on Image Processing*, vol. preprint, p. 1-14.

Chen, Yi-Chen, Vishal M. Patel, Sumit Shekhar, Rama Chellappa, and P. Jonathon Phillips. 2013. "Video-based Face Recognition via Joint Sparse Representation". In *Automatic Face and Gesture Recognition*. (Shangai, China 2013).

Cohen, Ira, Fabio G. Cozman, Nicu Sebe, Marcelo C. Cirelo, and Thomas S. Huang. 2004. "Semisupervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-Computer Interaction". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, n° 12, p. 1553-1567.

Connolly, J.-F., E. Granger, and R.230 Sabourin. March 2010a. "An Adaptive Classification System for Video-Based Face Recognition". *Information Sciences*, vol. 192, p. 50-70.

Connolly, Jean-François, Eric Granger, and Robert Sabourin. 2008. "Supervised incremental learning with the fuzzy ARTMAP neural network". In *Artificial Neural Networks in Pattern Recognition. Third IAPR Workshop, ANNPR 2008*. p. 66–77.

Connolly, Jean-Francois, Eric Granger, and Robert Sabourin. 2012. "Evolution of heterogeneous ensembles through dynamic particle swarm optimization for video-based face recognition". *Pattern Recognition*, vol. 45, n° 7, p. 2460 - 2477.

Connolly, J.F., E. Granger, and R. Sabourin. 2010b. "An Adaptive Ensemble of Fuzzy ARTMAP Neural Networks for Video-Based Face Classification". In *Proceedings on IEEE World Congress on Computational Intelligence*.

Dagan, I. and S.P. Engelson. July 9-12 1995. "Committee-based sampling for training probabilistic classifiers". In *Proc. I. Conf. on Machine Learning*. (Sn Francisco, USA 1995), p. 150-7.

Davis, Jesse and Mark Goadrich. 2006. "The relationship between precision-recall and ROC curves". In *ACM International Conference Proceeding Series*. (Pittsburgh, PA, United states 2006), p. 233 - 240.

De-la Torre, M, E. Granger, R. Sabourin, and D. O. Gorodnichy. December, 2013 2013. "An Individual-Specific Strategy for Management of Reference Data in Adaptive Ensembles for Face Re-Identification". In *5THh I. C. on Imaging for Crime Detection and Prevention (ICDP)*. (London, U. K. 2013), p. 1-7.

De-la Torre, Miguel, Eric Granger, Paulo V. W. Radtke, Robert Sabourin, and Dmitry O. Gorodnichy. June 2012a. "Incremental Update of Biometric Models in Face-Based Video Surveillance". In *Proc. IJCNN*. (Brisbane, Australia 2012), p. 1-8.

De-la Torre, Miguel, Paulo V. W. Radtke, Eric Granger, Robert Sabourin, and Dmitry O. Gorodnichy. July 2012b. "A Comparison of Adaptive Matchers for Screening of Faces in Video Surveillance". In *Symposium on Computational Intelligence for Security and Defence Applications*. (Ottawa, Canada 2012), p. 1-8.

De-la Torre, Miguel, Eric Granger, Paulo V.W. Radtke, Robert Sabourin, and Dmitry O. Gorodnichy. 2014a. "Partially-Supervised Learning from Facial Trajectories for Face Recognition in Video Surveillance". *Information Fusion*, vol. in press, p. 1-39.

De-la Torre, Miguel, Eric Granger, Robert Sabourin, and Dmitry O. Gorodnichy. 2014b. "An Adaptive Ensemble-Based System for Face Recognition in Person Re-Identification". *Machine Vision and Applications (Accepted with revision)*, vol. XX, p. 1-29.

Despiegel, V., S. Gentric, and JC. Fondeur. March 2012. "Border control: From Technical to Operational Evaluation". In *International Biometric Performance Testing Conference*. (Gaithersburg, Maryland, US 2012).

Dewan, M.A.A., E. Granger, F. Roli, R. Sabourin, and G.L.. Marciallis. December 2013. "A Comparison of Adaptive Appearance Methods for Tracking Faces in Video Surveillance". In *International Conference on Imaging for Crime Detection and Prevention (ICDP)*. (London, UK. 2013), p. 1-8.

Didaci, Luca and Fabio Roli. 2006. "Using co-training and self-training in semi-supervised multiple classifier systems". In *LNCS (including LNAI and LNB)*. (Hong Kong, China 2006), p. 522 - 530.

Diehl, C. P. and G. Cauwenberghs. 2003. "SVM Incremental learning, adaptation and optimization". In *Proceedings of the International Joint Conference on Neural Networks*. p. 2685-2690.

Ditzler, G., M.D. Muhlbaier, and R. Polikar. 2010. "Incremental Learning of New Classes in Unbalanced Datasets: Learn++.UDNC". In *Proceedings 9th International Workshop, Multiple Classifier Systems, MCS 2010*. (Berlin, Germany 2010), p. 33 - 42.

Ditzler, Gregory and Robi Polikar. July 2010. "An Ensemble Based Incremental Learning Framework for Concept Drift and Class Imbalance". In *WCCI 2010 IEEE World Congress on Computational Intelligence*.

Ditzler, Gregory and Robi Polikar. 2013. "Incremental Learning of Concept Drift from Streaming Imbalanced Data". *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, n° 10, p. 2283-2301.

Doddington, George, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. 1998. "Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance". In *International conference on spoken language processing*. p. 1351-1354.

Drummond, Chris. 2006. "Discriminative vs. generative classifiers for cost sensitive learning". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. (Quebec City, Que., Canada 2006), p. 479 - 490.

Drummond, Chris and Robert C. Holte. 2006. "Cost curves: An improved method for visualizing classifier performance". *Machine Learning*, vol. 65, n° 1, p. 95 - 130.

du Plessis, Marthinus Christoffel and Masashi Sugiyama. 2012. "Semi-Supervised Learning of Class Balance under Class-Prior Change by Distribution Matching". *CoRR*, vol. abs/1206.4677, p. 1-26.

Duda, Richard O., Peter E. Hart, and David G. Stork, 2001. *Pattern Classification*. ed. 2. Michigan, U.S. : Wiley.

Ekenel, Hazim Kemal, Lorant Szasz-Toth, and Rainer Stiefelhagen. 2009. "Open-set face recognition-based visitor interface system". In *LNCS*. (Liege, Belgium 2009), p. 43 - 52.

Ekenel, Hazim Kemal, Johannes Stallkamp, and Rainer Stiefelhagen. May 2010. "A video-based door monitoring system using local appearance-based face models". *Computer Vision Image Understanding*, vol. 114, n° 5, p. 596–608.

El Gayar, Neamat, Shaban A. Shaban, and Sayed Hamdy. 2006. "Face recognition with semi-supervised learning and multiple classifiers". In *Proc. of WSEAS Int. Conf. on Computational Intelligence, Man-Machine Systems and Cybernetics*. (USA 2006), p. 296–301.

Fan, Wei, Salvatore J. Stolfo, Junxin Zhang, and Philip K. Chan. 1999. "AdaCost: Misclassification Cost-Sensitive Boosting". In *Proceedings of the Sixteenth International Conference on Machine Learning*. (San Francisco, CA, USA 1999), p. 97–105. Morgan Kaufmann Publishers Inc.

Fawcett, T. 2004. "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers". *Intelligent Enterprise Technologies Laboratory*.

Fawcett, T. 2006. "An Introduction to ROC Analysis". *Pattern Recognition Letters*, vol. 27, n° 8, p. 861-874.

Fischer, Mika, HazımKemal Ekenel, and Rainer Stiefelhagen. 2011. "Person re-identification in TV series using robust face recognition and user feedback". *Multimedia Tools and Applications*, vol. 55, n° 1, p. 83-104.

Flach, Peter and Edson Matsubara. 2008. "On classification, ranking, and probability estimation". In *Probabilistic, Logical and Relational Learning - A Further Synthesis*. (Dagstuhl, Germany 2008), p. 1-10. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

Flach, Peter A. July 2004. "The many faces of ROC analysis in machine learning". Tutorial Slides, ICML'04.

Forman, G. 2008. "Quantifying counts and costs via classification". *Data Mining and Knowledge Discovery*, vol. 17, n° 2, p. 164 - 206.

Forman, George. 2006. "Quantifying trends accurately despite classifier error and class imbalance". (Philadelphia, PA, United states 2006), p. 157 - 166.

Franco, A., D. Maio, and D. Maltoni. 2010. "Incremental template updating for face recognition in home environments". *Pattern Recognition*, vol. 43, n° 8, p. 2891 - 903.

Freni, Biagio, Gian Luca Marcialis, and Fabio Roli. 2008. "Template Selection by Editing Algorithms: A Case Study in Face Recognition". In *Proc. Joint Int. Association of Pattern Recognition International*. (Orlando, USA 2008), p. 745-754.

Fritzke, B. 1996. "Growing self-organizing networks-why?". In *4th European Symposium on Artificial Neural Networks, ESANN '96. Proceedings*. (Brussels, Belgium 1996), p. 61 - 72.

Fu, L., H. H. Hsu, and J. C. Principe. 1996. "Incremental backpropagation learning networks". *IEEE Trans. on Neural Networks*, vol. 7, n° 3, p. 757-761.

Galar, Mikel, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2011. "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches". *IEEE Transactions on Systems, Man and Cybernetics*, vol. 42, p. 463-484.

Goh, Rodney, Lihao Liu, Xiaoming Liu, and Tsuhan Chen. 2005. The CMU Face In Action Database. *Analysis and Modelling of Faces and Gestures*, p. 255–263. Carnegie Mellon University.

Gomez-Sanchez, E., Y.A. Dimitriadis, J.M. Cano-Izquierdo, and J. Lopez-Coronado. jan 2002. "mu;ARTMAP: use of mutual information for category reduction in Fuzzy ARTMAP". *Neural Networks, IEEE Transactions on*, vol. 13, n° 1, p. 58 -69.

Gonzalez-Castro, Victor, Rocio Alaiz-Rodriguez, and Enrique Alegre. 2013. "Class distribution estimation based on the Hellinger distance". *Information Sciences*, vol. 218, p. 146-164.

González-Castro, Víctor, Rocío Alaiz-Rodríguez, Laura Fernández-Robles, R. Guzmán-Martínez, and Enrique Alegre. 2010. "Estimating Class Proportions in Boar Semen

Analysis Using the Hellinger Distance ". In *Proceedings of the 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems - Volume Part I*. (Berlin, Heidelberg 2010), p. 284–293. Springer-Verlag.

Gorodnichy, D.O. 2005. " Video-based framework for face recognition in video ". In *Proceedings. The 2nd Canadian Conference on Computer and Robot Vision*. (Piscataway, NJ, USA 2005), p. 330 - 8.

Gouaillier, Valerie. April 2009. *Intelligent video surveillance: Promises and challenges. technological and commercial intelligence report*. Technical report. Centre de Recherche Informatique de Montreal (CRIM) and Technopole Defence and Security.

Graham, J. and J.A. Starzyk. 2008. " A hybrid self-organizing neural gas based network ". In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. (Piscataway, NJ, USA 2008), p. 3806 - 13.

Granger, Eric, P. Henniges, Robert Sabourin, and L. S. Oliveira. 2007. " Supervised Learning of Fuzzy ARTMAP Neural NetworksThrough Particle Swarm Optimization ". *J. of Pattern Recognition Research*, vol. 2, p. 27-60.

Guo, Xinjian, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. 2008. " On the class imbalance problem ". In *2008 Fourth International Conference on Natural Computation*. (Piscaatway, NJ, USA 2008), p. 192 - 201.

Haker, Steven, William M. Wells III, Simon K. Warfield, Ion-Florin Talos, Jui G. Bhagwat, Daniel Goldberg-Zimring, Asim Mian, Lucila Ohno-Machado, and Kelly H. Zou. 2005. " Combining classifiers using their receiver operating characteristics and maximum likelihood estimation ". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. (Palm Springs, CA, United states 2005), p. 506 - 514.

Hampapur, A., L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti. March 2005. " Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking ". *IEEE Signal Processing Magazine*, vol. 22, n° 2, p. 38- 51.

Hart, P. may 1968. " The condensed nearest neighbor rule (Correspondence) ". *IEEE Transactions on Information Theory*, vol. 14, n° 3, p. 515 - 516.

Hewitt, Robin and Serge Belongie. 2006. " Active learning in face recognition: Using tracking to build a face model ". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. (New York, NY, United states 2006), p. 157.

Jain, Anil K. and Arun Ross. 2002. " Learning User-Specific Parameters in a Multibiometric System ". In *Int. Conf. on Image Processing*. p. 57-60.

Jervis, B.W., T. Garcia, and E.P. Giahnakis. 1999. " Probabilistic simplified fuzzy ARTMAP (PSFAM) ". *IEE Proceedings: Science, Measurement and Technology*, vol. 146, n° 4, p. 165 - 169.

Kachites McCallum, A. and K. Nigam. July 24-27 1998. “Employing EM and pool-based active learning for text classification”. In *Proc. I. Conf. on Machine Learning*. (San Francisco, USA 1998), p. 350-8.

Kadirkamanathan, V. and M. Niranjan. 1993. “A function estimation approach to sequential learning with neural networks”. *Neural Computation*, vol. 5, n° 6, p. 954 - 75.

Kamgar-Parsi, B., W. Lawson, and B. Kamgar-Parsi. 2011. “Toward Development of a Face Recognition System for Watchlist Surveillance”. *IEEE Trans. PAMI*, vol. 33, n° 10, p. 1925 - 37.

Kapp, M.N., C.Od.A. Freitas, and R. Sabourin. 2007. “Methodology for the design of NN-based month-word recognizers written on Brazilian bank checks”. *Image and Vision Computing*, vol. 25, n° 1, p. 40 - 9.

Khreich, W., E. Granger, A. Miri, and R. Sabourin. 2010a. “A Comparison of Techniques for On-line Incremental Learning of HMM parameters in Anomaly Detection”. In *Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications*. p. 2732 - 52.

Khreich, W., E. Granger, A. Miri, and R. Sabourin. 2010b. “Iterative Boolean Combination of classifiers in the ROC space: An application to anomaly detection with HMMs”. *Pattern Recognition*, vol. 43, n° 8, p. 2732 - 52.

Khreich, W., E. Granger, A. Miri, and R. Sabourin. July 2012. “Adaptive ROC-Based Ensemble of HMMs applied to Anomaly Detection”. *Pattern Recognition*, vol. 45, p. 208-230.

Kittler, J. 1998. “Combining Classifiers: A theoretical Framework”. *Pattern Analysis and Applications*, vol. 1, p. 18-27.

Klare, B. and A.K. Jain. Sept 2010. “On a taxonomy of facial features”. In *International Conference on Biometrics: Theory Applications and Systems (BTAS)*. p. 1-8.

Kubat, Miroslav and Stan Matwin. 1997. “Addressing the Curse of Imbalanced Training Sets: One-Sided Selection”. In *In Proceedings of the Fourteenth International Conference on Machine Learning*. p. 179–186. Morgan Kaufmann.

Kuncheva, L.I., 2004. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley.

Landgrebe, Thomas C. W. et al. 2006. “Precision-Recall Operating Characteristic (P-ROC) curves in imprecise environments”. In *Proceedings of ICPR*. p. 123 - 127.

Lee, Kuang-Chih, J. Ho, Ming-Hsuan Yang, and D. Kriegman. 2003. “Video-based face recognition using probabilistic appearance manifolds”. In *Proc. in CVPR*. p. I-313-I-320.

Lee, Kuang-Chih, J. Ho, Ming-Hsuan Yang, and D. Kriegman. 2005. " Visual tracking and recognition using probabilistic appearance manifolds ". *Computer Vision and Image Understanding*, vol. 99, p. 303-331.

Lee, Tsu-Chang. 1990. " Structure level adaptation for artificial neural networks: theory, applications, and implementations ". PhD thesis, Stanford University, Stanford, CA, USA. Adviser-Peterson, Allen M.

Lerner, B. and H. Guterman. 2008. " Advanced Developments and Applications of the Fuzzy ARTMAP Neural Networks in Pattern Classification ". *Techniques in Data Analysis and Applications in Advanced Computational Intelligence*, , p. 77-107.

Lewis, David D. and Jason Catlett. 1994. " Heterogeneous Uncertainty Sampling for Supervised Learning ". In *Proceedings of the Eleventh International Conference on Machine Learning*. p. 148–156. Morgan Kaufmann.

Li, Fayin and H. Wechsler. 2005. " Open set face recognition using transduction ". *IEEE Trans. on PAMI*, vol. 27, n° 11, p. 1686-97.

Li, Kai and Hou-Kuan Huang. 2002. " Incremental learning proximal support vector machine classifiers ". In *Proceedings of 2002 International Conference on Machine Learning and Cybernetics (Cat.No.02EX583)*. (Piscataway, NJ, USA 2002), p. 1635 - 7.

Lim, C. P. and R. F. Harrison. June 1995. " Probabilistic Fuzzy ARTMAP: An autonomous neural network architecture for bayesian probability estimation ". In *Artificial Neural Networks, 1995., Fourth Int. Conf. on*. p. 148-153.

Lim, C. P. and R. F. Harrison. 1997. " An incremental adaptive network for on-line supervised learning and probability estimation ". *Neural Networks*, vol. 10, n° 5, p. 925-939.

Liu, Weifeng, José C. Principe, and HaykinSimon, 2010. *Kernel Adaptive Filtering: A Comprehensive Introduction*. Wiley.

Liu, Xiaoming and Tsuhan Cheng. 2003. " Video-based face recognition using adaptive Hidden Markov Models ". In *Proceedings 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. (Los Alamitos, CA, USA 2003), p. 340 - 5.

Liu, Xu-Ying, Jianxin Wu, and Zhi-Hua Zhou. April 2009. " Exploratory Undersampling for Class-Imbalance Learning ". *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, n° 2, p. 539-550.

Lopez, Victoria, Alberto Fernandez, Salvador Garcia, Vasile Palade, and Francisco Herrera. 2013. " An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics ". *Information Sciences*, vol. 250, n° 0, p. 113 - 141.

Lu, Ke, Zhengming Ding, Jidong Zhao, and Yue Wu. 2010. " A novel semi-supervised face recognition for video ". In *Proc. of the International Conference on Intelligent Control and Information Processing*. p. 313-316.

Lu, Zhenyu, Xindong Wu, and J. Bongard. 2009. " Active learning with adaptive heterogeneous ensembles ". In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining (ICDM 2009)*. (Piscataway, NJ, USA 2009), p. 327 - 36.

Marcel, S. and G. Rodriguez, Y. Heusch. 2007. " On the Recent Use of Local Binary Patterns for Face Authentication ". *International Journal of Image and Video Processing - Special issue on Facial Image Processing*.

Marcialis, Gian Luca and Fabio Roli. 2002. " Fusion of LDA and PCA for Face Verification ". In *Proceedings of the Workshop on Biometric Authentication*.

Marcialis, G.L., A. Rattani, and F. Roli. 2008. " Biometric template update: an experimental investigation on the relationship between update errors and performance degradation in face verification ". In *Structural, Syntactic, and Statistical Pattern Recognition. Proceedings Joint IAPR International Workshop, SSPR & SPR 2008.* (Berlin, Germany 2008), p. 684-93.

Martinez, A.M. and A.C. Kak. 2001. " PCA versus LDA ". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, n° 2, p. 228-33.

Matta, F. and J.-L. Dugelay. 2009. " Person recognition using facial video information: a state of the art ". *Journal of Visual Languages and Computing*, vol. 20, n° 3, p. 180-7.

Matta, Federico and Jean-Luc Dugelay. 2006. " Video face recognition: A physiological and behavioural multimodal approach ". In *Proceedings - International Conference on Image Processing, ICIP*. (San Antonio, TX, United states 2006), p. IV497-IV500.

Merati, A., N. Poh, and J. Kittler. 2010. " Extracting discriminative information from cohort models ". In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*. p. 1 -6.

Mou, D., R. Schweer, and A. Rothermel. 2006. " A self-learning video-based face recognition system ". In *Proceedings of the International Conference on Consumer Electronics, digest of technical papers*. p. 97-8.

Mou, Dengpan, 2010. *Machine-based Intelligent Face Recognition*. Springer and Higher Education Press.

Muhlbaier, M., A. Topalis, and R. Polikar. June 2004. " Learn++.MT: A new approach to incremental learning ". *5th Int. Workshop on Multiple Classifier Systems (MCS 2004)*, vol. 3077, p. 52-61.

Muhlbaier, M. D., A. Topalis, and R. Polikar. 2009. "Learn++.NC: Combining Ensemble of Classifiers With Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes". *Neural Networks, IEEE Transactions on*, vol. 20, n° 1, p. 152-168.

Nickabadi, A., M.M. Ebadzadeh, and R. Safabakhsh. 2008. "Evaluating the performance of DNPSO in dynamic environments". In *2008 IEEE International Conference on Systems, Man and Cybernetics*. (Piscataway, NJ, USA 2008), p. 2640-5.

Oh, I. and C. I. Suen. 2002. "A class-modular feedforward neural network for handwritting recognition". *Pattern Recognition*, vol. 35, p. 229-244.

Oh, Sangyoon, Min Su Lee, and Byoung-Tak Zhang. 2011. "Ensemble Learning with Active Example Selection for Imbalanced Biomedical Data Classification". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, n° 2, p. 316-325.

Ojala, T., M. Pietikainen, and D. Harwood. 1996. "A comparative study of texture measures with classification based on feature distributions". *Pattern Recognition*, vol. 29, n° 1, p. 51-9.

Ojala, T., M. Pietikainen, and T. Maenpaa. 2002. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". *IEEE Tr. PAMI*, vol. 24, n° 7, p. 971-87.

Okada, K., L. Kite, and C. von der Malsburg. 2001. "An adaptive person recognition system". In *Proceedings 10th IEEE International Workshop on Robot and Human Interactive Communication*. (Piscataway, NJ, USA 2001), p. 436-41.

Osorio, F. S. and B. Amy. 1999. "Inss: A hybrid system for constructive machine learning". *Neurocomputing*, , p. 191-205.

Ozawa, S., Soon Lee Toh, S. Abe, Shaoning Pang, and N. Kasabov. 2005. "Incremental learning of feature space and classifier for face recognition". *Neural Networks*, vol. 18, n° 5-6, p. 575 - 84.

Pagano, Christophe, Eric Granger, Robert Sabourin, and Dmitry O. Gorodnichy. June 2012. "Detector Ensembles for Face Recognition in Video Surveillance". In *IJCNN*. (Brisbane, Australia 2012), p. 1-8.

Platt, John. 1991. "A resource-allocating network for function interpolation". *Neural Comput.*, vol. 3, n° 2, p. 213–225.

Poh, Norman, Rita Wong, Josef Kittler, and Fabio Roli. 2009. "Challenges and research directions for adaptive biometric recognition systems". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. (Alghero, Italy 2009), p. 753-764.

Polikar, R. 2006. “Ensemble based systems in decision making”. *IEEE Circuits and Systems Magazine*, vol. 6, n° 3, p. 21 - 44.

Polikar, R., L. Udpa, S. S. Udpa, and V. Honavar. 2001. “Learn++: An Incremental Learning Algorithm for MLP Networks”. *IEEE Trans. SMC*, vol. 31, n° 4, p. 497-508.

Polikar, R., J. Byorick, S. Krause, A. Marino, and M. Moreton. 2002. “Learn++: a classifier independent incremental learning algorithm for supervised neural networks”. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN’02 (Cat. No.02CH37290)*. (Piscataway, NJ, USA 2002), p. 1742 - 7.

Polikar, Robi, Joseph DePasquale, Hussein Syed Mohammed, Gavin Brown, and Ludmilla I. Kuncheva. 2010. “Learn++.MF: A random subspace approach for the missing feature problem”. *Pattern Recognition*, vol. 43, p. 3817-3832.

Radtke, Paulo, Eric Granger, Robert Sabourin, and Dmitry Gorodnichy. 2013a. Adaptive ensemble selection for face re-identification under class imbalance. Zhou, Z.-H., Fabio Roli, and Josef Kittler, editors, *Multiple Classifier Systems*, volume 7872 of *Lecture Notes in Computer Science*, p. 95-108. Springer Berlin Heidelberg. ISBN 978-3-642-38066-2.

Radtke, Paulo V.W., Eric Granger, Robert Sabourin, and Dmitry O. Gorodnichy. 2013b. “Skew-sensitive boolean combination for adaptive ensembles – An application to face recognition in video surveillance”. *Information Fusion*, vol. 20, p. 31-48.

Rattani, A., G.L. Marcialis, and F. Roli. 2008a. “Capturing large intra-class variations of biometric data by template co-updating”. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. (Piscataway, NJ, USA 2008), p. 1-6.

Rattani, A., G.L. Marcialis, and F. Roli. 2008b. “Biometric template update using the graph mincut algorithm : a case study in face verification”. In *Proceedings of the Biometrics Symposium*. (Piscaatway, NJ, USA 2008), p. 23-8.

Rattani, A., G.L. Marcialis, and F. Roli. 2009a. “An experimental analysis of the relationship between biometric template update and the Doddington’s Zoo: a case study in face verification”. In *Proceedings of the 15th International Conference on Image Analysis and Processing*. (Berlin, Germany 2009), p. 434-42.

Rattani, Ajita. 2010. “Adaptive Biometric System based on Template Update Procedures”. PhD thesis, University of Cagliari.

Rattani, Ajita, Biagio Freni, Gian Luca Marcialis, and Fabio Roli. 2009b. “Template update methods in adaptive biometric systems: A critical review”. In *Lecture Notes in Computer Science (included Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. (Alghero, Italy 2009), p. 847 - 856.

Roli, F., G. Fumera, and J. Kittler. 2002. "Fixed and trained combiners for fusion of imbalanced pattern classifiers". In *Information Fusion, 2002. Proceedings of the Fifth International Conference on*. p. 278 - 84.

Roli, F., L. Didaci, and G.L. Marcialis. August 2007. "Template co-update in multimodal biometric systems". In *International Conference on Biometrics*. (Seoul, Korea 2007), p. 1194 - 202.

Roli, Fabio and Gian Luca Marcialis. August 2006. "Semi-supervised PCA-based face recognition using self-training". In *JIAPR - Int. Workshop on Structural and Syntactical Pat. Rec. and Statistical Techniques in Pat. Rec.* (Hong Kong, China 2006), p. 560-568. Springer.

Roli, Fabio, Luca Didaci, and Gian Luca Marcialis. 2008. Adaptive biometric systems that can improve with use. Govindaraju, N. R. V., editor, *Advances in Biometrics: Sensors, Systems and Algorithms*, p. 447-471. Springer.

Ross, David A., Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. 2008. "Incremental Learning for Robust Visual Tracking". *Int. Journal of Computer Vision, Special issue: Learning for vision*, vol. 77, p. 125-141.

Ruping, S. 2001. "Incremental Learning with Support Vector Machines". In *IEEE International Conference on Data Mining*. p. 641-642.

Salmeron, M., J. Ortega, and C.G. Puntonet. 1999. "On-line optimization of radial basis function networks with orthogonal techniques". In *Foundations and Tools for Neural Modeling. International Work-Conference on Artificial and Natural Neural Networks, IWANN'99. Proceedings, Vol.1 (Lecture Notes in Computer Science Vol.1606)*. (Berlin, Germany 1999), p. 467 - 77.

Salmeron, Moises, Julio Ortega, Carlos G. Puntonet, and Alberto Prieto. 2001. "Improved RAN sequential prediction using orthogonal techniques". *Neurocomputing*, vol. 41, p. 153 - 172.

Satta, Riccardo. July 2013. "Appearance Descriptors for Person Re-Identification: A Comprehensive Review". arXiv:1307.5748v1.

Scheffer, T., C. Decomain, and S. Wrobel. 2001. "Active Hidden Markov models for information extraction". In *Proc. Int. Conf. in Advances in Intelligent Data Analysis*. (Berlin, Germany 2001), p. 309-18.

Scott, M.J.J., M. Niranjan, and R.W. Prager. 1998. "Realisable classifiers: improving operating performance on variable cost problems". In *BMVC 98. Proceedings of the Ninth British Machine Vision Conference*. (Southampton, UK 1998), p. 306 - 15.

Shannon, C. E. 1948. "A Mathematical Theory of Communication". *The Bell Systems Technical Journal*, vol. 27, p. 379-423, 623-656.

Singh, Richa, Mayank Vatsa, Arun Ross, and Afzel Noore. 2010. "Biometric Classifier update using online learning: A case study in near infrared face verification". *Image and Vision Computing*, vol. 28, p. 1098-1105.

Syed, N. A., H. Liu, and K. K. Sung. 1999. "Incremental Learning with Support Vector Machines". In *Proceedings of International Conference on Artificial Intelligence*.

Tang, E.K., P.N. Suganthan, and X. Yao. 2006. "An analysis of diversity measures". *Machine Learning*, vol. 65, n° 1, p. 247 - 71.

Tao, Qian and R. Veldhuis. 2008. "Hybrid fusion for biometrics: combining score-level and decision-level fusion". In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*. (Piscataway, NJ, USA 2008), p. 1 - 6.

Tao, Qian and R. Veldhuis. 2009. "Threshold-optimized decision-level fusion and its application to biometrics". *Pattern Recognition*, vol. 42, n° 5, p. 823 - 36.

Tax, D.M.J. and R.P.W. Duin. 2008. "Growing a multi-class classifier with a reject option". *Pattern Recognition*, vol. 29, n° 10, p. 1565 - 70.

Tomek, I. nov. 1976. "Two Modifications of CNN". *IEEE Trans. on Systems, Man and Cybernetics*, vol. SMC-6, n° 11, p. 769 -772.

Turk, M. and A. Pentland. January 1991. "Eigenfaces for Recognition". *Journal of Cognitive Neuroscience*, vol. 3, n° 1, p. 71-86.

Verzi, S.J., G.L. Heileman, M. Georgiopoulos, and M.J. Healy. 4-8 1998. "Boosted ARTMAP". In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*. p. 396 -401 vol.1.

Vigdor, B. and Lerner B. November 2007. "The Bayesian ARTMAP". *IEEE Transactions on Neural Networks*, vol. 18, n° 6.

Viola, P. and M. Jones. 2004. "Robust real-time face detection". *International Journal of Computer Vision*, vol. 2, n° 57, p. 137-154.

Wang, E. H. and A. Kuh. 1992. "A smart algorithm for incremental learning". *In International Joint Conference on Neural Networks*, vol. 3, p. 121-126.

Wang, Huafeng, Yunhong Wang, and Yuan Cao. 2009. "Video-based face recognition: A survey". *Proceedings of World Academy of Science, Engineering and Technology*, vol. 60, p. 293 - 302.

Wang, Shuo and Xin Yao. 2009. "Diversity analysis on imbalanced data sets by using ensemble models". In *CIDM'09*. p. 324-331.

Wang, Shuo, L.L. Minku, D. Ghezzi, D. Caltabiano, P. Tino, and Xin Yao. Aug 2013a. "Concept drift detection for online class imbalance learning". In *Neural Networks (IJCNN), The 2013 International Joint Conference on*. p. 1-10.

Wang, Shuo, L.L. Minku, and Xin Yao. April 2013b. "A learning framework for online class imbalance learning". In *Computational Intelligence and Ensemble Learning (CIEL), 2013 IEEE Symposium on*. p. 36-45.

Williamson, James R. 1996. "Gaussian ARTMAP: A Neural Network for fast incremental-learning of noisy Multidimensional Maps". *Neural Networks*, vol. 9, p. 881-897.

Wu, Fangjun. November 2012. "Comparing Boosting and Cost-Sensitive Boosting With Imbalanced Data". *Journal of Convergence Information Technology*, vol. 7, n° 21, p. 1-8.

Yang, M-H., D. J. Kriegman, and N. Ahuja. January 2002. "Detecting Faces in Images: A Survey". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, n° 1, p. 34-58.

Yilmaz, Alper, Omar Javed, and Mubarak Shah. December 2006. "Object tracking: A survey". *ACM Comput. Surv.*, vol. 38, n° 4.

Yingwei, Lu, N. Sundararajan, and P Saratchandran. 1997. "A sequential learning scheme for Function approximation using minimal radial basis function Neural Networks". *Neural Computation*, vol. 9, p. 461-478.

Yu, Guoxian, Guoji Zhang, Carlotta Domeniconi, Zhiwen Yu, and Jane YouZ. 2012. "Semi-supervised classification based on random subspace dimensionality reduction". *Pattern Recognition*, vol. 45, n° 3, p. 1119 - 1135.

Zenobi, Gabriele and Pádraig Cunningham. 2001. Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. Raedt, L. and Peter Flach, editors, *Machine Learning: ECML 2001*, volume 2167 of *Lecture Notes in Computer Science*, p. 576-587. Springer Berlin Heidelberg. ISBN 978-3-540-42536-6.

Zhang, Cha and Zhengyou Zhang. June 2010. *A survey of recent advances in face detection*. Technical Report MSR-TR-2010-66. Microsoft Research.

Zhang, Y. and A. M. Martınez. 2006. "A weighted probabilistic approach to face recognition from multiple images and video sequences". *Image and Vision Computing*, vol. 24, p. 626-638.

Zhang, Yongbin and Aleix Martinez. 2004. "From Stills to Video: Face Recognition Using a Probabilistic Approach". In *Computer Vision and Pattern Recognition Workshop Conference on*. p. 78.

Zhao, W. et al. December 2003. “Face Recognition: A Literature Survey”. *ACM Computing Surveys*, vol. 35, n° 4, p. 399-458.

Zhou, S., V. Krueger, and Rama Chellappa. 2003. “Probabilistic recognition of human faces from video”. *Computer Vision and Image Understanding*, vol. 91, n° 1-2, p. 214 - 45.

Zhou, S.K., R. Chellappa, and B. Moghaddam. 2004. “Visual tracking and recognition using appearance-adaptive models in particle filters”. *Image Processing, IEEE Transactions on*, vol. 13, n° 11, p. 1491-1506.

Zhou, Z.-H. and Daoqiang Zhang. 2005. “(2D)2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition”. *Neurocomputing*, vol. 69, n° 1-3, p. 224-31.